Mauro M. Monsalve Mercado

Space in the Brain

OF LEARNING AND REPRESENTATIONS

Dissertation der
Graduate School of Systemic Neurosciences der
Ludwig-Maximilians-Universität München
Date of Submission: 9th of April 2018

Supervisor
Prof. Dr. Christian Leibold

Second reviewer
Prof. Dr. Andreas Herz

Defended on the 10th of July, 2018

Contents

PART I INTRODUCTION

Where is space in the brain?

The hippocampal formation 16

Place cells 18

Grid cells 20

Models of grid cell activity

Continuous attractor networks 24

Feedforward networks 27

Oscillatory interference model 29

PART II RESULTS

On spatial memory

On the representation of space

On grid cell firing patterns

PART III DISCUSSION

A reality check

How well can CANs path-integrate? 89

What else can we learn from OIMs? 91

What are experiments saying? 93

An alternative: Why feed-forward models? 95

Evidence for feed-forward models 97

What our model brings to the table 100

Outlook

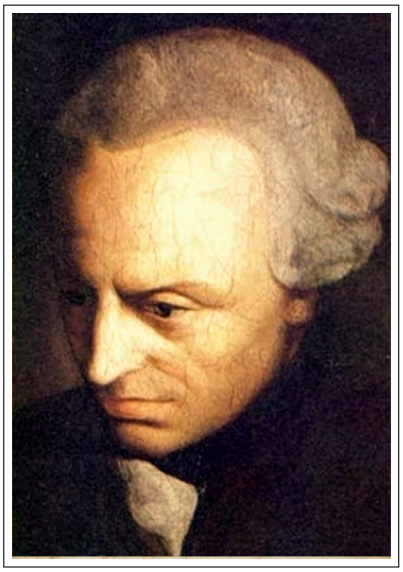
Para mi familia.

Overview

The debate about the nature and essence of space has entertained philosophers since early accounts dating back to ancient Greece. At the centre of the dispute is the issue of the methods of existence of space, putting into doubt if space is a proper entity by itself, exists as a relation among entities, or is just a conceptual framework for the human experience. For instance, Isaac Newton thought of space as absolute, existing independent of the rest of the world, while its contemporary Gottfried Leibniz viewed space as a set of geometrically defined relationships between objects, existing only as an idealised abstract construction. Immanuel Kant contrasted these two viewpoints and came to the conclusion that space existed in neither form but rather was imposed by peoples minds as a conceptual framework in which to organise the human experience. Early modern mathematicians and physicists (Gauss, Poincaré, Einstein) have formalised the concept of space and time within a scientific framework, where experimental observations and not just reasoning have the last word.

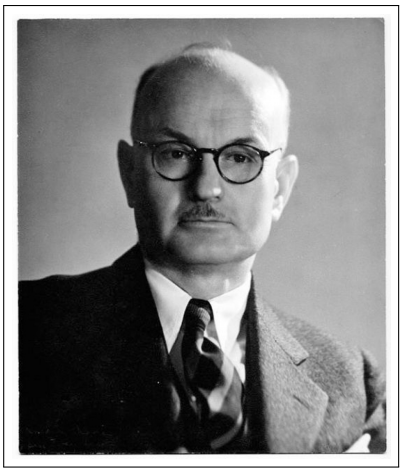
The debate is today most relevant in the field of neuroscience. There is an important distinction to be made between space in the physical world, as envisioned by modern physics, and peoples' perception of space, largely shaped by the human experience. That both perspectives are usually intermingled in philosophical discussions of the reality of space and time is only indicative of just how recently neuro-scientific studies on the brain's representation of space have weighted in on the discussion. Edward Tolman first hinted at the existence of a cognitive map of space in the brain, that is a representation of where objects are in the physical world and the spatial relationships among them.

In this thesis we are concerned with the question of how space is represented in the brain. What are the organisational features the brain is using to know where we are and where we are going? How is this organisation different from the conceptual frameworks developed to understand the physical world? The answers to these questions are both intuitive and complex, yet the field remains still in its infancy, with more questions than answers appearing every day.



Is space real?

In his Inaugural Dissertation of 1770 Kant presents his views on the nature and relation of space and mind: "Space is not something objective and real, nor a substance, nor an accident, nor a relation; instead, it is subjective and ideal, and originates from the mind's nature in accord with a stable law as a scheme, as it were, for coordinating everything sensed externally". Kant wonders about the existence of space, and argues that perhaps space is only realised in peoples' minds. 18th century portrait by unidentified painter from Wikimedia Commons.



Is space in the brain?

In his work, Tolman (1948) makes an early reference to the development of spatial maps in the brain: "We believe that in the course of learning, something like a field map of the environment gets established in the rat's brain... The stimuli ... are usually worked over ... into a tentative, cognitive-like map of the environment. And it is this tentative map, indicating routes and paths and environmental relationships, which finally determines what responses, if any, the animal will finally release." Portrait from Wikimedia Commons. Citation selected by O'Keefe (2014) for his Nobel lecture.

In this work, particularly large efforts are dedicated to understand the patterns of activity of the so-called grid cells in the medial entorhinal cortex of mammals. The importance of these cells is two-fold.

First, theoretical and experimental arguments suggest they have a prominent role in path integration, the mechanism underlying an animals' ability to self-actualise its location with respect to local cues by means of self-motion and internal information. In addition, some involvement of these cells in goal-directed navigation and efficient coding of contextual associations to space merit a more profound investigation into the patterns of activity and related coding schemes implemented by these subnetworks.

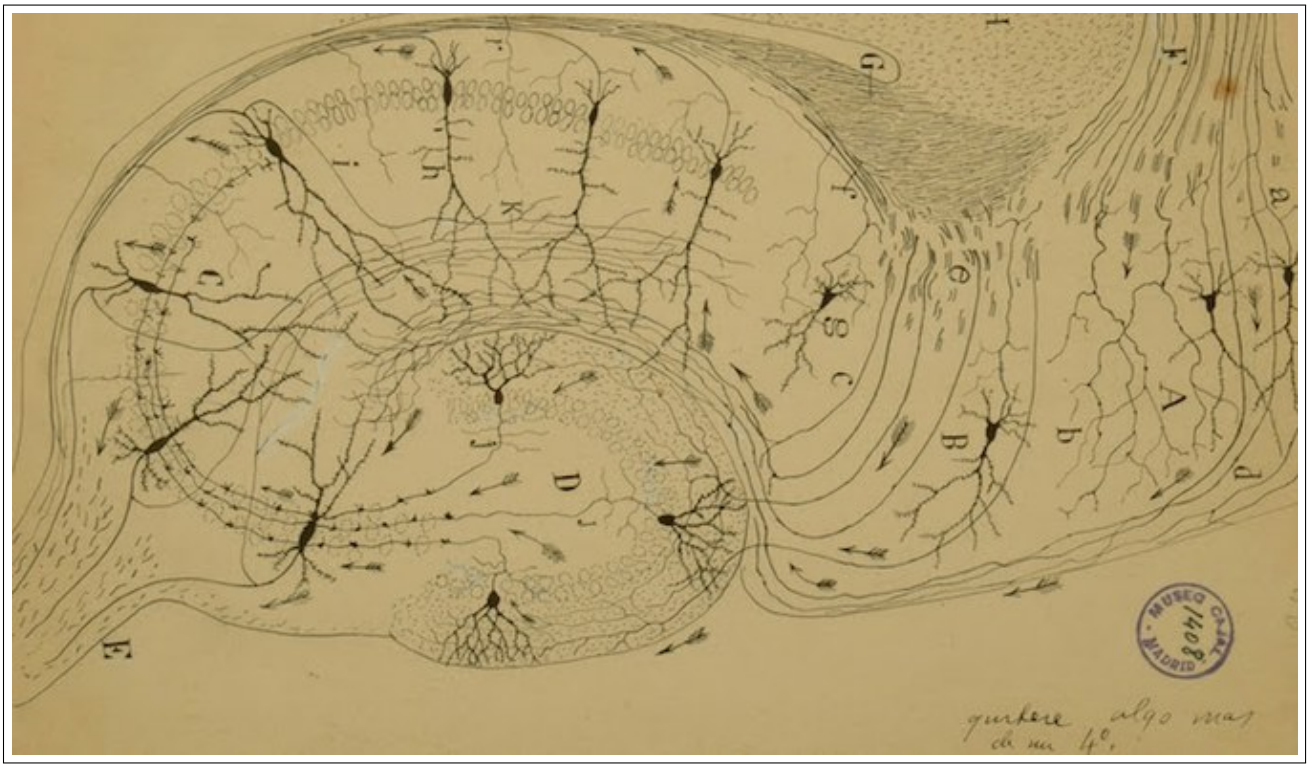
A second, more general point, is related to the key position wherein grid cells are found. The entorhinal cortex, whose importance was already mentioned by Ramón y Cajal more than a century ago, is the bridge that connects high-end cortical areas to the hippocampus. On the one hand, high-end cortices are characterised by their multimodal representation of rather abstract encoded information, and thus the study of how they fulfil their functional role can be challenging. On the other hand, even though far from the receiving end of sensory information, the hippocampus is well-known for their clear-cut spatially modulated cells, with well-defined correlates to a fundamental attribute of the environment. As an in-between step, the entorhinal cortex can help us to elucidate some of the fundamental coding principles and information processing mechanisms employed by high-end cortices, all the while retaining easy to probe neural correlates to spatial navigation.

As an introduction to the problem, I'll review some of the most important experimental observations about how space is represented in the brain. For this, I discuss the anatomical organisation of the hippocampal information, the brain region responsible for hosting the neural correlates of space. Within this region, most cells involved in encoding information about spatial navigation can be found, including place cells and grid cells. A special review of these two cell types will help us notice similarities and differences highlighting what makes grid cells unique. Existing computational models try to capture the main features of grid cells' patterns of activity. An introductory review of these models is accompanied by a comprehensive discussion at the end of the manuscript, where special focus is given to the contrast between the explained experimental features against the unexplained observations.

Several peer-reviewed publications show in detail my contributions to understanding the representation of space. Of particular interest is the computational model of grid cell activity resolving some of the most puzzling observations still unexplained.

PART I

INTRODUCTION



An early drawing of the structure and neural circuitry of the rodent hippocampal formation. By Santiago Ramón y Cajal. Histologie du Systeme Nerveux de l'Homme et des Vertebretes, Vols. 1 and 2. A. Maloine. Paris. 1911. Ramon y Cajal © Cajal Institute, CSIC. Madrid, Spain.

Where is space in the brain?

The question is phrased to incite a double meaning. On the one hand the question makes reference to a particular region of the brain, if any, completely in charge of making us understand where we are and where we are going. Fortunately, we've know for some time that a structure within the medial temporal lobe, namely the hippocampal formation, plays a pivotal role in the encoding of spatial information and episodic memory. Impairment to this region results in severe memory loss and reduces performance in spatial navigation tasks.

Having a closer look at what neurons in this region are doing answers an alternative interpretation of our original question. Where do abstract constructions of space lie in the brain? or in more concrete terms, where do we find neural correlates of space? The question was answered by O'Keefe and Dostrovsky (1971) who recorded cells in the hippocampus that fire whenever an animal walks over a specific region of its environment. Since then, more cells have been found in the hippocampal formation with correlates related to spatial navigation, such as head direction cells (Taube, Muller, and Ranck 1990), boundary cells (Barry et al. 2006), speed cells (Kropff et al. 2015), and grid cells (Hafting et al. 2005).

Cells functionally involved in spatial representation are found scattered all over the hippocampal formation, with subnetworks preferentially hosting specific functional types. How do all these cells influence each other? and what does this mean for the representation of space? A clue is to be found by paying attention to the neural pathways interconnecting the different subnetworks, which points us in the right direction if we are to understand the mechanism behind higher levels of information processing leading to the brain's ability of representing space.

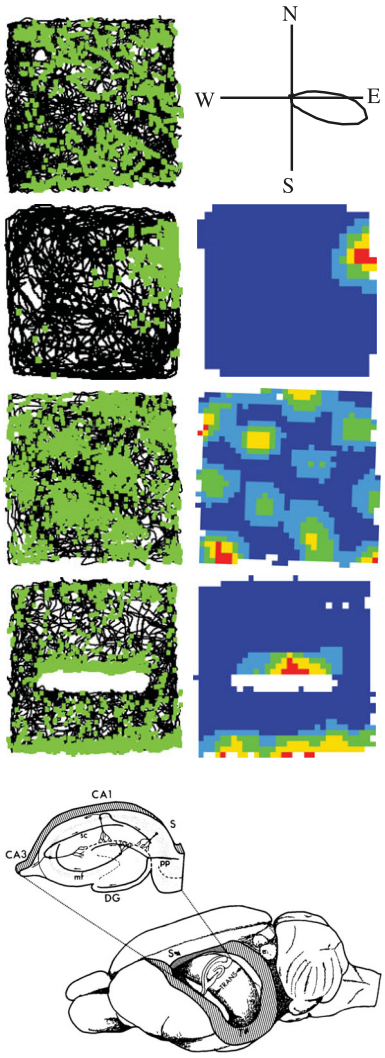


Figure 1: Where is space in the brain? *Space is represented by cells in the hippocampal formation (bottom: a rat's brain showing the hippocampus and a transversal slice identifying its subfields). Head direction cells (top first) fire when the animal's head points to a particular direction, and are found in presubiculum and EC (and many additional areas outside the HF). Place cells (top second) are active in localized areas and are prevalent in the hippocampus, typically in areas CA1 to CA3. Grid cells are a similar cell type but fire at multiple fields regularly spaced from each other, and are mostly found in EC (but can be found in pre- and parasubiculum). Another type, the border cell, fires close to physical boundaries, has been recorded in subiculum, pre- and parasubiculum and EC. Modified with permission from Hartley et al. 2014; Amaral and Witter 1989.*

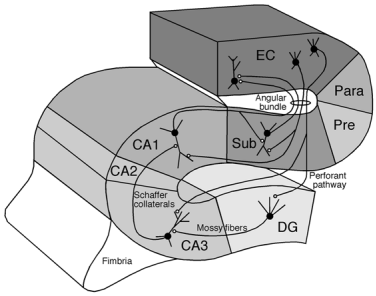


Figure 2: Hippocampal formation. *The diagram indicates major connectivity pathways among hippocampal subfields and the parahippocampal region. The hippocampus includes the dentate gyrus (DG) and subfields CA1-CA3 (sometimes called hippocampus proper). Contiguously are defined the areas subiculum, presubiculum (PrS), parasubiculum (PaS), and the entorhinal cortex (EC). The standard view on connectivity considers a monosynaptic pathway without reciprocal synapses. The major projections compose the trisynaptic loop, it includes the perforant path (EC to DG), the mossy fibers (DG to CA3), and the Schaffer collaterals (CA3 to CA1). Area CA1 and subiculum close the loop by projecting back to the entorhinal cortex via the angular bundle. Modified with permission from Andersen et al. 2006.*

The hippocampal formation

The hippocampus and adjacent areas are structures located in the medial temporal lobe of mammalian brains, although equivalent structures can be identified in birds and reptiles, perhaps playing a similar role in spatial navigation and memory. In rodents, the hippocampus occupies a rather large portion of the brain, as an evident elongated C-shaped structure going rostrally from the septal nuclei to the ventrocaudal temporal cortex (see Figure 1). The dorsoventral axis (also called septotemporal or longitudinal axis) follows the elongated dimension of the hippocampus, with transversal slices usually displaying most subfields depending on the position along the axis (see Figure 2). The axis has a significantly unequal cortical and subcortical input/output profile, with spatially modulated cells found in more dorsocaudal locations.

Different criteria can tell us where to set different brain areas apart. Anatomical location, cyto- and chemoarchitecture, and connectivity studies can all help to distinguish relevant subfields with functional implications for behaviour and information processing. For instance, the entorhinal cortex plays a key role as the bridge between abstract high-end cortices and more sub-cortically defined hippocampal fields. It raises the question of its involvement in the representation of space from a biological point of view, in terms of what intrinsic unique properties are responsible for the observed activity in the hippocampal formation.

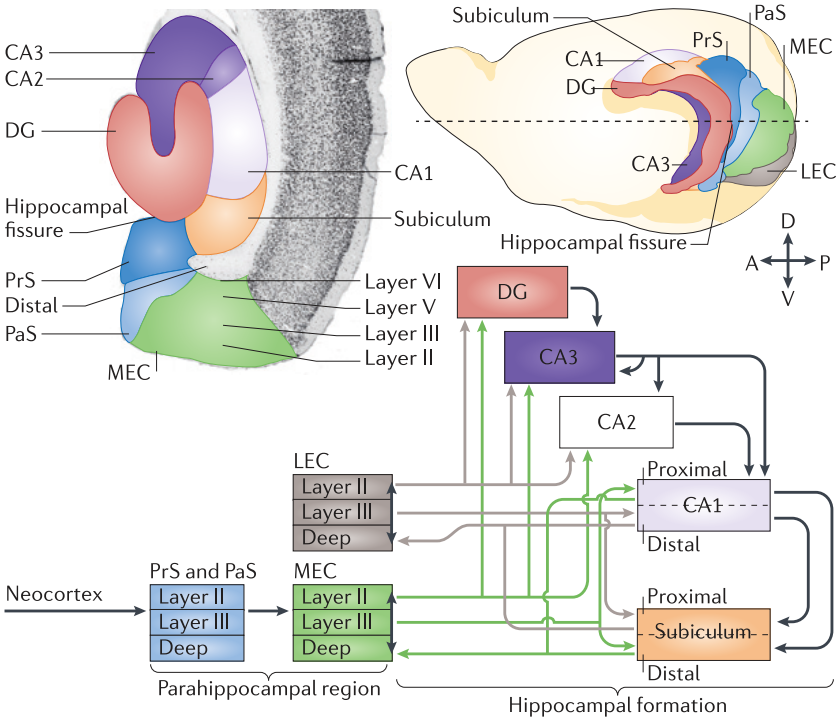


Figure 3: Anatomical connectivity in the rodent hippocampal formation. *A closer look reveals a more intricate connectivity layout. Different anatomical locations within subfields might target separate areas or even distinct parts of the same area. For instance, the back projection from hippocampus to EC terminates in deeper layers of the cortical laminar structure, while superficial layers are the main origin of the perforant pathway targeting hippocampus. Modified with permission from Moser et al. 2014.*

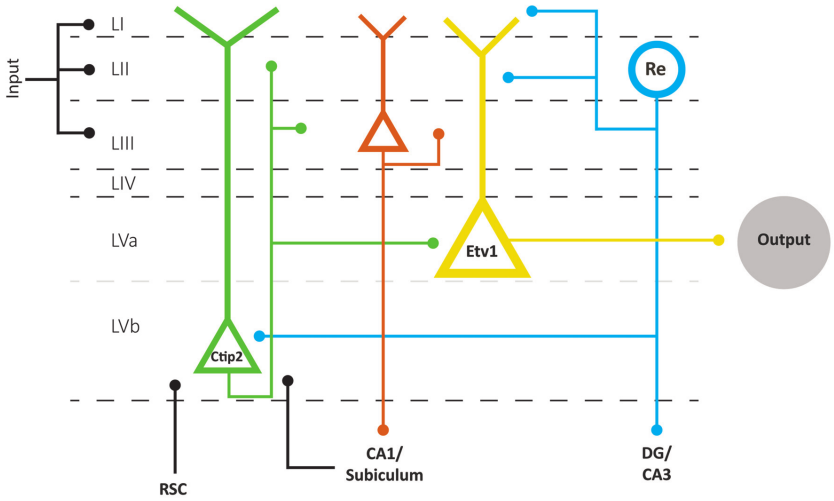


Figure 4: Entorhinal cortex circuitry. *An additional level of complexity can be found within entorhinal cortex. Principal cells with varied physiological and morphological properties dominate the different layers and exhibit unique connectivity patterns. Deep layer Vb pyramidal cells receive information from intrahippocampal fields CA1 and subiculum, and in turn relay it forward to superficial layers II and III. Stelate cells (or rather reelin positive cells) in layer II process this information and send it back to cells in layer Vb and hippocampal areas DG and CA3. Modified from Witter et al. 2017 under open access license CC BY.*

In the context of spatially modulated cells connectivity layouts are of fundamental importance. Grid cells recorded in the entorhinal cortex beg us to examine its communication pathways with the hippocampus (see Figure 3 for details). Further understanding can be gain by examining the complex intrinsic connectivity within EC itself (see Figure 4). In particular, deeper layers of EC receive strong projections from spatial cells in CA1 and subiculum, which after processing are returned back to DG and CA3. An even more complicated picture includes an extra level of organisation in layer II (see Figure 5). Two types of principal cells organise each in separated subnetworks, wherein principal cells communicate only via various kinds of inhibitory interneurons. Subnetworks may in addition interact by means of intermediate excitatory cell types, increasing the difficulty in understanding the spatial selectivity of grid cells.

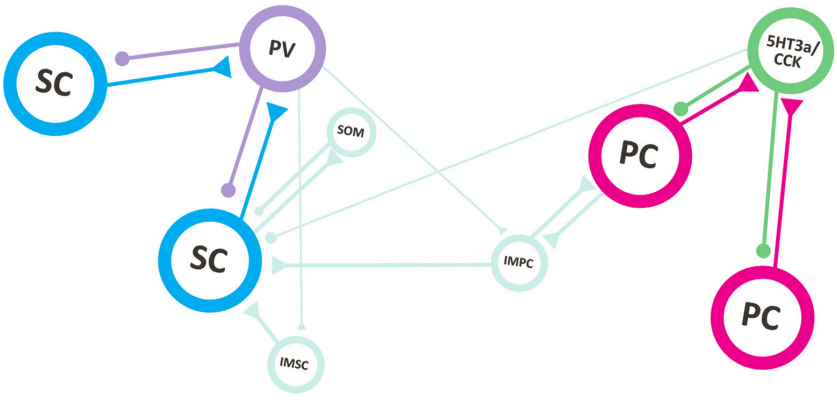


Figure 5: Entorhinal cortex Layer II. *Even at the level of a single layer we can find non-trivial functional networks. Pyramidal and stellate cells in layer II form separate interconnected networks through specialized inhibitory interneurons. Neither type of principal cell presents direct excitatory connections between them or the other type, only communicating through interneurons. Stellate cells communicate via fast-spiking parvalbumin positive interneurons in a many-to-one and back to one-to-many fashion, while pyramidal cells talk through 5HT3a/CCK interneurons. Both subnetworks may interact by way of physiological and morphological identifiable intermediate types of stellate and pyramidal cells. Modified from Witter et al. 2017 under open access license CC BY.*

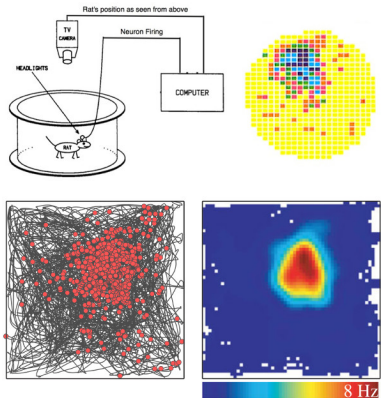


Figure 6: Place cell recordings *can tell us how the hippocampus encodes spatial locations. The animals movements (grey trace) are tracked via an LED on its head, registering the locations where the cell fires a spike (red dots). The activity is typically visualised with a firing rate map (histogram), highlighting areas of increased activity. Modified with permission from Muller, Kubie, and Ranck 1987; Jeewajee et al. 2014.*

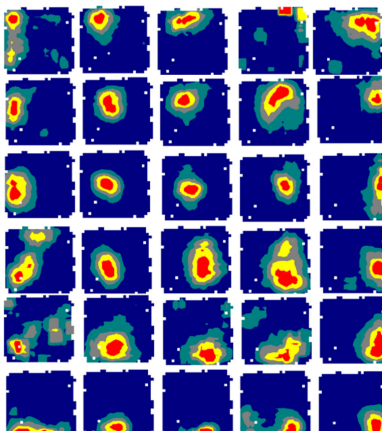


Figure 7: Place cells as a cognitive map. *A population of place cells is able to represent all locations within the environment using a firing rate code. As the animal explores its surroundings, a sequence of active cells conveys information about its location. Modified from O'Keefe 2014.*

Place cells

The earliest work directly linking the hippocampus involvement in encoding an animals environment dates back decades to its first electrophysiological recordings (O'Keefe and Dostrovsky 1971; O'Keefe 1976; O'Keefe and Nadel 1978). The firing activity of principal cells was found to be restricted to confined areas within the environment, suggesting that they may underlie the neural substrate for the brains' cognitive map of space.

One fundamental requirement of a spatial cognitive map is the neuronal representation of all locations in an environment. Figure 7 shows an example of how a population of place cells seem to form a complete representation of a square box environment, where different cells fire for distinct but overlapping areas of the box. A given location then elicits heterogeneous firing from a subset of place cells, and the highly redundant code ensures robustness of the representation.

A population of cells smoothly changes its firing profile while sampling locations in the box, suggesting a decoding mechanism based on the cells' instantaneous firing rate. The firing rate code must be stable in a given environment over long periods of time. It has been shown in fact that place cells can retain their spatial correlates over a period of at least several months (Thompson and Best 1990).

Another essential qualification for a cognitive map is the ability to tell apart dissimilar environments from each other. Substantial alterations to a given environment induce arbitrary changes to the cells firing maps (e.g. adjusted rates, map relocation, no firing at all), resulting in a new population code uniquely identifying the novel contextual situation (e.g colours, shape of the enclosure, salient cues). The effect that environmental or contextual changes has on firing maps is studied under the umbrella term 'remapping' (Muller and Kubie 1987).

The effect of contextual information is not all equal along the dorsoventral axis of the rodent hippocampus (Moita et al. 2003; Jung, Wiener, and McNaughton 1994; Maurer et al. 2005; Kjelstrup et al. 2008; Royer et al. 2010). Ventral cells are more susceptible to emotional cues (e.g. fear, stress, rewards), have lower spatial information, stability, and theta modulation. As a consequence, a lower proportion of ventral cells qualify for representing space, and those remaining are less accurate than their dorsal counterparts. One important related finding is the increase in place field size along the dorsoventral axis, although no agreement has been reached on whether the increment comes in rather gradual or discrete steps (but some hints come from the functional organisation of the axis, see Strange et al. 2014).

Temporal coding

A firing rate code is not the only way to establish a cognitive map of space. There is ample evidence suggesting an additional information layer encoded by means of further temporal modulation of place cell activity[1]. O'Keefe and Recce (1993) and Skaggs et al. (1996) first brought to attention the tendency of place cells to fire bursts of spiking activity at regular intervals in the theta band. When compared to the simultaneously recorded local field potential (LFP), it was evident that the periodicity of place cell firing was just slightly faster, appearing to precess with respect to the LFP signal (see Figure 8). From a coding perspective, this phase precession phenomenon reproduces the firing sequence of a set of cells with overlapping place fields, with the added advantage of broadcasting this relationship within each theta cycle, a much faster timescale than that offered by a firing rate code (Dragoi and Buzsáki 2006). Its usefulness is evident in learning. For instance, temporal encoding of place cell sequences within theta timescales can explain the swift formation of grid cells firing patterns in the entorhinal cortex undergoing spike-time dependent plasticity (Monsalve-Mercado and Leibold 2017).

Huxter, Burgess, and O'Keefe (2003) discuss evidence in favour of a dual coding mechanism in the activity of place cells. They argue that the instantaneous firing rate and the precise temporal firing may independently encode information for distinct variables such as location within a field or speed of movement. Typically, enough freedom appears to be present for the dual code to be involved in the hippocampus role in spatial and episodic memory.

For the most part, experiments dealing with temporal coding focus on linear track foraging, leaving open the question up to what extent are these results true in the open field. Jeewajee et al. (2014) and Huxter et al. (2008) investigate what behavioural correlates are the best predictors of the firing activity in open field foraging. They found strong phase precession to be best correlated with measures of spatial exploration, such as distance run through the field, trajectory, and heading, among other proxy measures of space.

[1] O'Keefe and Recce, Hippocampus, 1993: "Phase relationship between hippocampal place units and the EEG theta rhythm".

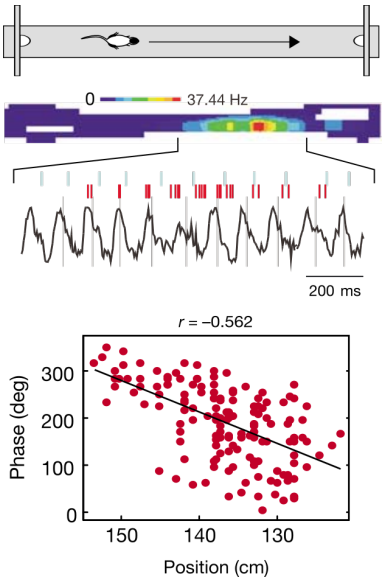


Figure 8: On phase precession. *Place cells spiking activity is in addition temporally modulated. Spike bursts (red ticks) happen always at earlier phases with respect to the local EEG theta rhythm (black trace). A single cell firing pooled over runs shows that the phases of theta cycles at which spikes happen have a linear relation to position on the track (bottom plot). Modified with permission from Huxter, Burgess, and O'Keefe 2003.*

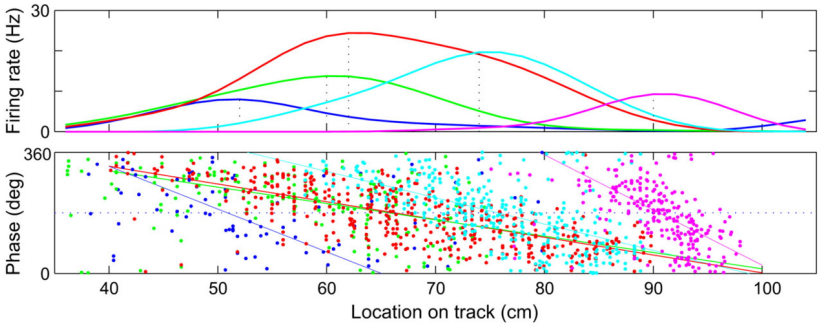


Figure 9: Temporal encoding of spatial information. *A cells periodicity of firing within the field is tuned such that the total range of phase precession completes roughly one cycle. As a result, cells with different place field sizes will fire at corresponding periodicities in order to carry out a cycle of phase precession. Modified with permission from Dragoi and Buzsáki 2006.*

Grid cells

When Brun et al. (2002) removed all input from hippocampal area CA3 to CA1, they observed that CA1 place cells still displayed robust place fields. It pointed towards a remaining strong projection from the entorhinal cortex as the main alternative source of spatial information. The first recordings in superficial layers of medial entorhinal cortex unveiled the existence of cells with multiple spatially receptive fields, whose hallmark feature, being roughly equally spaced from each other, justified the label of 'grid cells' [2] (Fyhn et al. 2004; Hafting et al. 2005), see Figure 10 for details.

Grid cells' firing maps have distinct geometrical qualities (see Figure 11). For maps with enough regularity, quantitative measures define the grid scale (average spacing between neighbouring fields) and the grid orientation and phase (location of fields) relative to other cells' maps or local landmarks (usually the recording enclosure). An important observation is that neighbouring cells react consistently to environmental changes. For instance, relative orientations and phases within these networks are maintained across different environments (Fyhn et al. 2007).

These subnetworks have been observed in all anatomical positions along the dorsoventral axis (Stensola et al. 2012). They constitute a discrete organization of the representation into independent modules of coherent cell activity. This coherence property of the grid cell representation system suggests networks might self-organise, since they have been found to independently respond to changes in the environment, individually adjusting their orientation and phases, and sharing similar pattern irregularities and theta-frequency modulation within networks. Perhaps their most distinctive attribute is the tendency to exhibit a larger grid spacing towards more ventral locations along the longitudinal axis, in much the same way as hippocampal place cells do, although different scales can coexist in overlapping anatomical locations (see Figure 12).

Grid fields regularity and the MEC's key position in the hippocampal formation first led to the conjecture that grid cells may have a

[2] Hafting et al., Nature, 2005: "Microstructure of a spatial map in the entorhinal cortex."

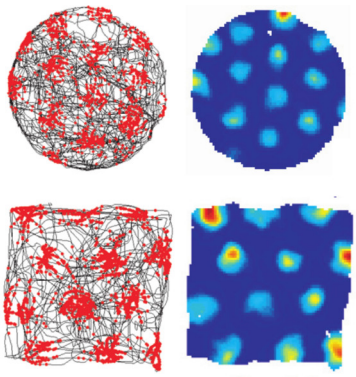


Figure 10: Grid cells receptive fields. *Electrophysiological recordings of principal cells in superficial layers of MEC revealed regularly spaced multiple receptive fields. The regularity of the pattern strongly depends on the level of symmetry of the recording enclosure. Grid cells have been found in rats, mice, bats, monkeys, and humans, suggesting it might be widely conserved through mammalian species. Modified with permission from Hafting et al. 2005.*

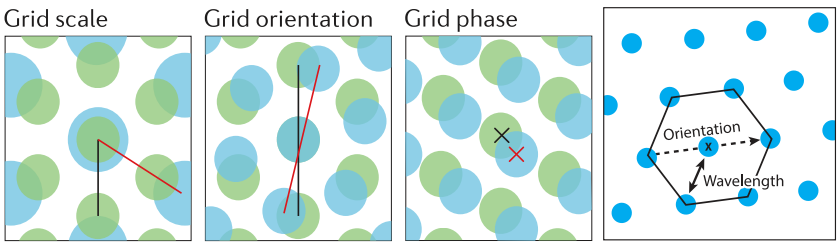


Figure 11: Grid cells basic properties. *The cells firing patterns can be characterised in terms of their geometrical configurations. Assuming a regular enough triangular arrangement, it is possible to measure the average spacing between fields (grid scale), and the relative grid orientation and grid phase (location of fields) with respect to other cells patterns or local cues (e.g the enclosure). Modified with permission from Moser et al. 2014.*

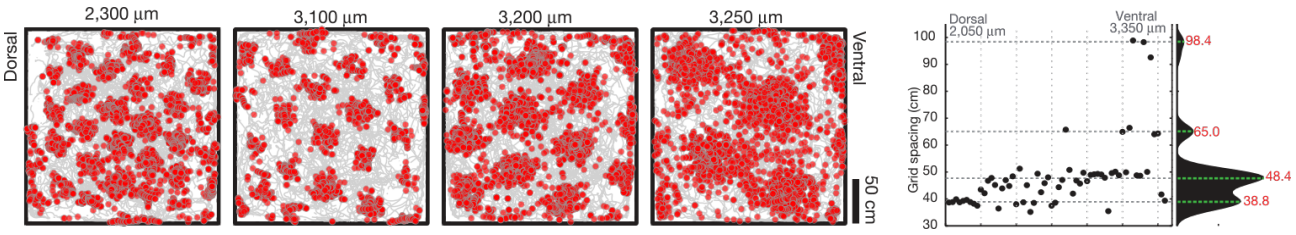


Figure 12: Grid cells cluster in modules. *Recordings along the dorsoventral axis of MEC reveal that grid cells are organised in independent modules. Cells belonging to a local module have firing maps sharing similar geometrical properties, such as spacing, orientation, and field size, but that in general evenly cover the whole range of possible grid phases. High density recordings uncover a tendency for modules to have a larger spacing for more ventrally located positions (first four panel are labelled by their distance to the postrhinal border), although modules with various spacings can be found in anatomically overlapping locations (see rightmost panel). Modified with permission from Stensola et al. 2012.*

central role in path integration, since unlike place cells, substantial contextual changes are only able to shift or rotate the patterns, highlighting a prominent influence of self-motion cues on maintaining the stability of patterns.

Further motivation to study grid cells comes from theoretical considerations. It has been conjectured that provided certain read-out mechanisms, grid cells unique properties constitute an optimal neural substrate to represent an animals location and navigate space to the best degree of accuracy with minimal expenditure (Mathis, Herz, and Stemmler 2012b; Mathis, Herz, and Stemmler 2012a; Mathis, Stemmler, and Herz 2015; Stemmler, Mathis, and Herz 2015).

A challenge to grid cells primary role as the brains' metric system arises from the patterns irregularities, since some of their coding principles may strongly rely on a perfectly triangular arrangement. The patterns, however, are far from ideal. They are only roughly triangular, present strong field to field variability (Dunn et al. 2017), and are affected by external perceptual inputs such as novelty (Barry et al. 2007; Barry et al. 2012). In addition, the geometry of the recording enclosure (level of symmetry, smooth deformations), and local cues can cause noticeable distortions in the pattern and influence orientation, local spacing, and field size (see Figure 13).

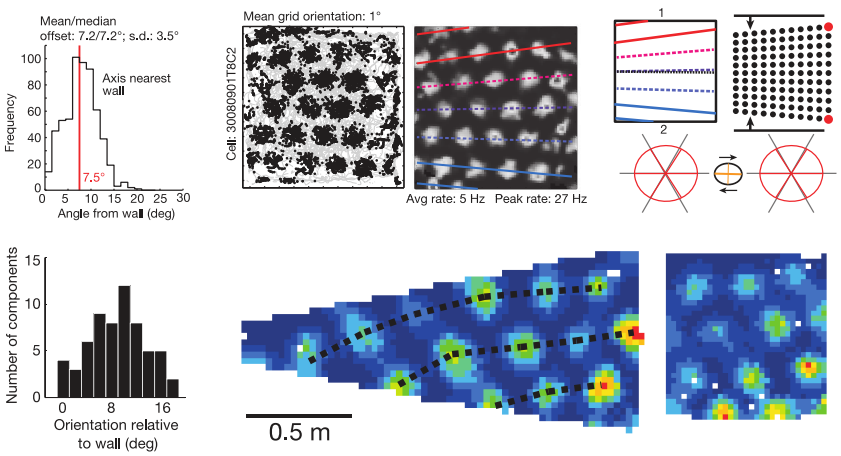


Figure 13: Distortions due to the arena. *Environment-induced distortions make grid cells less likely to represent the context-independent metric for navigation. Reports show that the patterns are biased to certain orientations to the walls of polarized enclosures (such as squares, but not circles, see first column panels from two independent studies). Furthermore, the maps appear to be rather localized, creating irregularities such as orientation gradients, ellipticity, shearing, bending, and so forth. Highly polarised environments such as an acute trapezoid (against a square recording of the same cell in the figure) might strongly influence the pattern to the point of negatively impacting regularity. Modified with permission from Stensola et al. 2015; Krupic et al. 2015.*

Models of grid cell activity

What drives us to understand grid cells' firing maps? One of the motivations behind a large body of work attempting to explain the appearance of the regularly spaced grid fields is their key role in spatial navigation and episodic memory. But beyond understanding the brain's way of representing space, the regular maps of grid cells offer an attractive example from the perspective of modelling dynamical systems.

The appearance of this kind of periodic arrangement of activity is not at all uncommon in nature. A prominent example in biology is the emergence of spots, stripes, spirals and similar patterns in the skin (or surface) of a broad range of animals. Turing (1952) first characterised the phenomena as a consequence of the self-organising activity of competing chemicals in a reaction-diffusion system. Further work by Gierer and Meinhardt (1972) made the concept broader to any biological system exhibiting competition as short-range excitation and long-range inhibition. The concept was first brought to the realm of neural systems by Amari (1977), who showed how stable patterns of neural activity of the Turing type could be implemented by networks of interconnected neurons in the presence of effective Mexican-hat connectivity.

Most successful grid cell models developed so far are of the Turing type. In the following I review the most recent advancements concerning mechanistic explanations for the emergence of hexagonal activity patterns, focusing on a perspective that emphasizes the underlying Turing mechanism and the assumptions behind the different biological implementations used to that end.

Continuous attractor networks

O'Keefe and Burgess (2005) first proposed the idea of an attractor network in the dMEC as a plausible explanation for grid field formation. The proposal argues in favour of the grid cell system as the brain's path integrator, an idea previously introduced by McNaughton et al. (1996) and Samsonovich and McNaughton (1997) in the context of the place cell system in area CA3. The periodicity of grid fields, however, allows for more optimal path integration throughout the entire environment. It generalises the one-dimensional attractor model for the head direction system introduced by Zhang (1996), with the added difficulties inherent to two-dimensional space, such as space not being a naturally periodic variable.

One of the first attempts at modelling the origin of hexagonality in the firing patterns of grid cells was done by Fuhs and Touretzky (2006)[3]. They build upon a framework originally proposed by Hopfield (1982) resembling a spin glass type neural network. In this type of networks model neurons are symmetrically connected to each other and interact only locally with its immediate neighbours. Under this configuration the system can dynamically reach a number of different stable states or attractors depending on the specific neuronal interactions in the network. In continuous attractor network models the symmetries of the resulting periodic patterns allow for equally favourable states to form a continuous manifold.

[3] Fuhs and Touretzky, The Journal of neuroscience, 2006: "A spin glass model of path integration in rat medial entorhinal cortex."

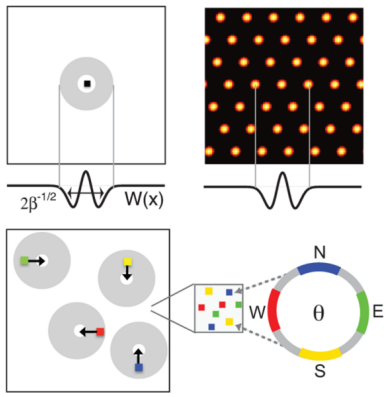


Figure 14: The attractor network model.
A neuron has a mexican-hat connectivity profile with its neighbours (top left). The network dynamics leads to a stable hexagonal configuration in the neural sheet (top right). All neurons are weakly biased towards one of the cardinal directions (bottom). Modified with permission from Burak and Fiete 2009.

A hexagonal pattern of activity can arise as the final steady state on an abstract two-dimensional sheet of non-linear neurons. The most important ingredient towards reaching a periodic stable state is the shape of the local connectivity between neurons. In continuous attractor models the connectivity profile is assumed to ultimately induce local short-range excitation and longer-range inhibition, commonly referred to as a Mexican-hat profile.

The top left part of figure 14 illustrates this connectivity profile for a central neuron in a square-lattice neuronal sheet. Such two-dimensional network configurations lead to the appearance of dynamically rich spatio-temporal periodic patterns (Murray 2004). The selection of the final stable pattern depends strongly on the non-linearity used to achieve stability. For continuous attractor models the commonly used threshold-linear rectification ensures the selection of the hexagonal pattern. The top right panel shows an example of the final state of activity in the neural sheet, where it is highlighted how the periodicity of the pattern in the neural sheet depends on the range of interaction of the Mexican-hat profile.

The model so far described explains the appearance of a static hexagonal configuration of activity bumps in an abstract neural sheet

of neurons. How can this be related to the actual firing patterns of grid cells as seen during exploration of the environment? For the pattern to appear in real two-dimensional space the activity of a neuron in the sheet has to follow the movement of the animal in space. Conklin and Eliasmith (2005) have shown that translating the activity bumps to track the animal's position implements correct path integration on attractor networks.

The models are capable of performing path integration by introducing a small bias towards a particular preferred direction in the neural sheet. The bias is included both in the connectivity profile and directly in the dynamics of the neurons' firing activity. The former is illustrated in the bottom part of figure 14, where each cell's connectivity profile is slightly shifted in one of the four cardinal directions (the cell's preferred direction). Additionally, the firing rate dynamics of each cell is coupled to the animal's running speed and directional signals, as it favours an increment in the firing rate of cells whose preferred direction is close to the current running direction. The overall effect forces the entire activity pattern to translate in the current direction of the animal's movement, thus achieving path integration to a certain degree of accuracy.

The translation of the population activity pattern in the neural sheet is illustrated in figure 15. From left to right it is depicted the development in time of the neural sheet (timestamps at the bottom). The top and bottom panels follow networks with periodic and aperiodic-damped boundary conditions respectively, showing that in the neural sheet both connectivity schemes are able to translate the pattern in concert with the animal's movement.

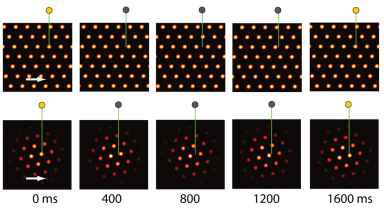


Figure 15: Path integration on CANs *is achieved with a high degree of accuracy for both periodic boundary conditions (top) and aperiodic ones (bottom). The bumps of activity in the neural sheet are translated in response to the animal's speed signal. Modified with permission from Burak and Fiete 2009.*

Computational model

Fuhs and Touretzky considered in their model a set of recurrently connected dMEC neurons. They model the dynamics of a neuron's membrane potential ξ_i using the voltage-based model developed by Hopfield (1984) with an additional driving force v_i coupling the velocity signal to the membrane dynamics

$$\tau \frac{d\xi_i}{dt} = -\xi_i + \sum_j W_{ij} \sqrt{\xi_j} \Theta(\xi_j) + v_i + \varepsilon, \tag{1}$$

where τ is the membrane integration time constant, Θ denotes the Heaviside function and expresses the neuron's transfer function converting voltage to firing rate, and ε is just some random Gaussian noise to break the initial symmetry and drive the pattern formation process.

The connectivity matrix W is composed by symmetric and asymmetric parts added together. The symmetric weight matrix comes

from the assumed Mexican-hat profile[4] and is responsible for selecting a periodic pattern. The asymmetric part is modelled by offsetting an inhibitory bump of connectivity from the centre of the cell in the neural sheet, and it's partly responsible for helping move the population activity along the animal's physical trajectory.

Finally, the pattern is translated via the action of the velocity coupling v_i. It ensures errors in path integration remain small as long as the membrane potential dynamics can catch up with rapid changes in speed. The coupling v_i grows linearly with the animal's speed but it's tuned to a preferred direction in a quality similar to that of head direction cells. It is in fact presumed that this tuning is related to prominent head direction input found in layers III and V of dMEC.

[4] In the original paper, the actual connectivity profile resembles a quickly decaying oscillation. It is learned via Hebbian plasticity during development assuming the propagation of randomly oriented wave packets of activity in the dMEC (see below). As long as the profile selects for a particular wavelength the pattern formation process will lead to hexagonality.

The continuous attractor model relies in a specific neural connectivity inherent to dMEC to develop spatial patterning. However, direct measurements of internal connectivity are challenging in nature, and strong evidence in favour of CANs is still lacking. Another alternative class of models, termed feedforward networks, achieves hexagonal pattern formation by the same means as continuous attractors, that is by implementing a spatio-temporal Turing instability. The manner in which this implementation is done, however, is fundamentally different. In their original proposal, O'Keefe and Burgess suggest the origin of the attractor dynamics to be the result of interference patterns of entorhinal oscillatory activity (ultimately shaping the connectivity in dMEC via plastic changes). This oscillatory interference model relies on the organisation of spiking activity with respect to the characteristic hippocampal-entorhinal oscillatory rhythms. The next two sections offer a concise review of both classes of models.

Feedforward networks

A second class of models discusses whether the contribution of strong projections coming from hippocampal areas into deeper layers of entorhinal cortex might be in turn responsible for the emergence of the grid pattern. It suggests that hexagonality in grid cells arises as a self-organization process of its inputs through Hebbian plasticity.

Kropff and Treves (2008)[5] propose cell adaptation to be the mechanism behind the self-organization process. In their model, a population of grid cells receive input from a population of spatially modulated cells (e.g. place cells) that cover the entire environment (see Figure 16). As the animal begins to slowly explore its surroundings, each cell samples a combination of the input distribution weighted by each of the inputs synaptic strength into that cell. The cell's total input slowly increases or decreases in magnitude, and in response the cell undergoes adaptation to adjust accordingly its sensitivity to elicit an action potential.

A key idea throughout the model is the separation of timescales. The adaptation response answers to fast dynamics that drive the emergence of hexagonality within behavioural timescales of the order of seconds. However, adaptation alone can not hold the grid pattern for longer times and spatial stability is lost as consequence. An additional component of the model helps fix the pattern in place. Slow learning dynamics continuously changes the strength of synaptic connectivity of the incoming inputs emulating Hebbian plasticity. Initially, areas in the environment with slightly stronger total input develop the first bumps of activity, which are then slowly rearranged by the interplay of adaptation and plasticity within minutes of exploration. The final pattern tends toward hexagonality as the optimal arrangement reconciling adaptive pressures and long term stability.

In addition, grid cells in the network interact with each other by imposing constraints on the adaptation dynamics in order to approximately satisfy competition criteria. An unchanging average criterion prevents the population activity from tending towards increasingly higher levels, a common outcome in unconstrained Hebbian learning models. A second criterion, keeping a constant population sparseness, helps to drive the dynamics away from a winner-take-it-all regime.

The model produces stable hexagonal patterns in a robust manner against highly variable conditions. At first the input is modelled after idealised place cell activity, with each place field a small Gaussian bump of activity, and with the whole population evenly covering the environment. However, hexagonality is likewise obtained when each place field is modelled as more widely distributed activity (e.g.

[5] Kropff and Treves, Hippocampus, 2008: "The emergence of grid cells: Intelligent design or just adaptation?"

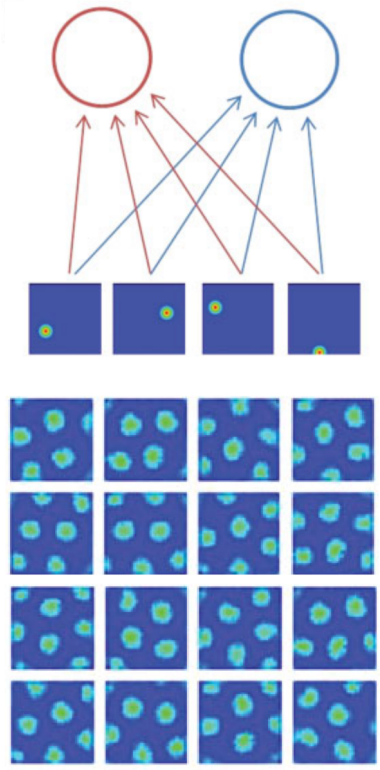


Figure 16: Feed-forward input model. *Each cell receives feed-forward input from a population of place cells (top). Adaptation dynamics drive the formation of hexagonal fields (bottom), a configuration slowly learned via Hebbian plasticity. Modified with permission from Kropff and Treves 2008.*

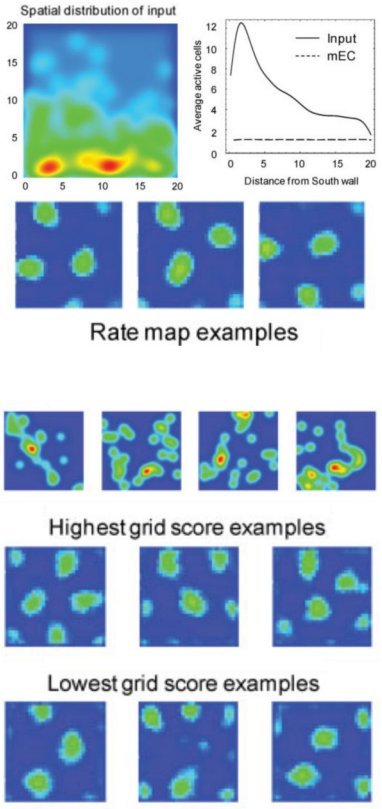


Figure 17: Model robustness. *Top: The density of inputs is higher closer to the south wall. Example rate maps show average grid cell activity and hexagonality remain stable. Similar performance measures are likewise obtained for widespread distributed input at the single cell level (bottom). Modified with permission from Kropff and Treves 2008.*

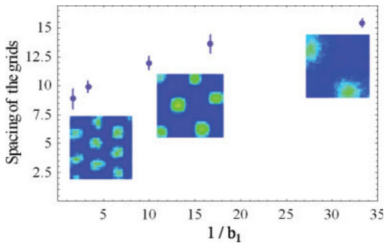


Figure 18: Spacing and adaptation. *The grid field spacing increases monotonically with the temporal constant of the activation variable. In all simulations the ratio $b_2 = b_1/3$ is kept fixed. Modified with permission from Kropff and Treves 2008.*

the sum of 20 widespread Gaussians), and is similarly successful for a less evenly distributed population input (e.g. a density of input fields four times higher closer to a particular wall that its opposite). An additional test of robustness compares learning performance of constant speed exploration against realistic speed statistics. Even for such high levels of speed variance, the pattern remains stable among learning and testing phases on random walk trajectories.

Computational model

Kropff and Treves (2008) considered a population of N_I place cells with firing rate r_j^t at time t projecting to N_{MEC} entorhinal neurons with synaptic weights J_{ij}, such that the total activity into neuron i

$$h_i^t = \frac{1}{N_I} \sum_{j=1}^{N_I} J_{ij} r_j^t \tag{2}$$

is a slowly changing spatially modulated combination of the inputs. To model cell adaptation, the slow input h is filtered through fast acting fatigue variables

$$r_{inact}^{t+1} = r_{inact}^t + b_2(h^t - r_{inact}^t) \tag{3}$$

$$r_{act}^{t+1} = r_{act}^t + b_1(h^t - r_{inact}^t - r_{act}^t) \tag{4}$$

with characteristic time constants $1/b_1, 1/b_2$ determining their interdependent dynamics. In general, $b_1 > b_2$ enables correct fatigue dynamics to develop and ensures the emergence of grid fields. After intermediate fatigue, the total input (currently $h = r_{act}(h)$) passes through a transfer function

$$\psi(h) = \psi_{sat} \frac{2}{\pi} arctan[g(h-\theta)]\Theta[h-\theta], \tag{5}$$

which finds the most optimal population threshold θ and gain g for that time step such that the average activity and sparseness

$$a = \frac{1}{N_{MEC}} \sum \psi_k, \quad s = N_{MEC}\, a^2 \sum \psi_k^2 \tag{6}$$

remain as unchanged as possible through the fatigue process. Finally, changes in activity are slowly written in the connectivity matrix

$$J_{ij}^{t+1} = J_{ij}^t + \varepsilon(\psi_i r_j - \langle \psi_i \rangle \langle r_j \rangle) \tag{7}$$

via a self-correcting Hebbian learning rule. In addition, weights are clipped at zero to maintain excitatory inputs only, and further normalized to keep the total input weight to any neuron constant.

Oscillatory interference model

The models discussed so far employ coding principles that are two dimensional in nature. Indeed, the appearance of a Turing instability able to drive hexagonal pattern formation is only possible in a two dimensional setting. A wildly different class of models attempts to address the periodicity of the grid pattern at its core by making it a superposition of purely periodic signals (Burgess, Barry, and O'Keefe 2007[6], Burgess 2008). It was formulated at first as an explanation of the phase precession phenomena observed in hippocampal places cells, however its periodic nature made it more suitable to model grid cell firing patterns in entorhinal cortex, which likewise exhibit spike phase precession.

The main idea of oscillatory interference models is remarkably simple. Stellate cells in entorhinal cortex receive external oscillatory inputs modulated by the entorhinal rhythms. One type of input just reflects the main component found in the local field potential and is modelled by a single oscillation with a baseline frequency in the theta band. A different kind of input is additionally modulated by internal motion signals, making its oscillatory frequency slightly higher than the baseline frequency. Together, the two types of input create an interference pattern reflected in a spatially periodic structure.

The second kind of input has been termed velocity controlled oscillators (VCOs) as an abstraction that models a broad range of possible biological implementations. It represents an oscillatory drive whose frequency growths linearly with the animal's speed and is additionally tuned to a preferred direction of movement. The interference pattern with the baseline theta oscillation produces a plane-wave of activity in the preferred direction of the VCO (first column in Figure 19). It is then argued that hexagonal patterns are the result of aggregated input from two or more VCOs with preferred directions that differ by an amount proportional to sixty degrees (see Figure 19).

An interference pattern consists of a fast oscillation modulated by a slower envelope. The fast component oscillates with the average frequency coming from the baseline and VCO oscillations, thus it is just slightly higher that baseline (theta) and naturally reproduces phase precession of spiking activity. On the other hand, the slow envelope's characteristic frequency is proportional to the difference of the input frequencies, and the resulting long wavelength determines the periodicity of the hexagonal pattern.

More precisely, the computational model assumes that the cell receives inputs whose active (time-dependent) frequencies $f_a(t)$ differ from a system-wide reference baseline frequency $f_b(t)$ as

$$f_a(t) = f_b(t) + \beta s(t) \cos(\phi(t) - \phi_d), \tag{8}$$

[6] Burgess, Barry, and O'Keefe, Hippocampus, 2007: "An oscillatory interference model of grid cell firing".

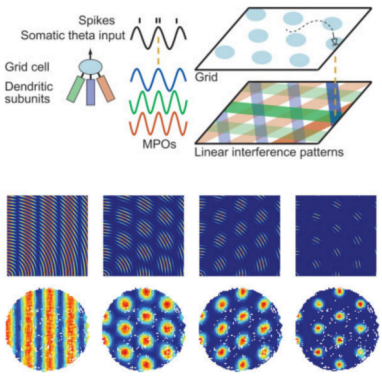


Figure 19: Oscillatory interference model. *Several membrane potential oscillators (MPOs) active on the target's cell dendrites combine in the soma to produce and interference pattern. A MPO is modulated by speed and head direction signals producing a planewave-like pattern in space (first column). It is then assumed two or more MPOs 60° apart are selected by an unspecified mechanism. Modified with permission from Burgess, Barry, and O'Keefe 2007.*

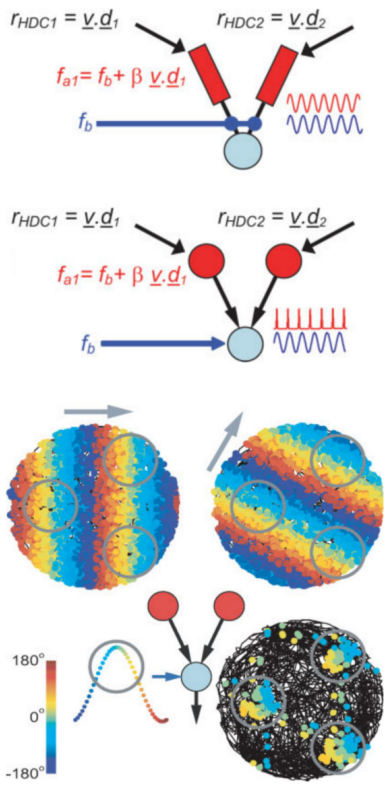


Figure 20: Biological implementations of the oscillatory interference model. *Velocity controlled oscillators (VCOs) can be implemented as dendritic membrane potential oscillations within the target cell or as synaptic input from motion-modulated spiking cells (top figure). Each VCO integrates motion in a preferred direction resulting in a spatial pattern formed of parallel stripes (bottom figure). Two or more VCOs with preferred directions 60° apart combine to form a hexagonal pattern. The resultant phase encodes spatial location via path integration. Modified with permission from Burgess 2008.*

where s, ϕ are the animal's speed and direction, ϕ_d is the input's preferred direction, and β is a positive constant reflecting the gain for the motion's path integration. The phase difference between the two oscillators at any given time $\varphi_{ab}(t)$ obeys $d\varphi_{ab}(t) = f_a(t) - f_b(t) = \beta s(t) \cos(\phi(t) - \phi_d)$, thus it integrates the motion of the animal in a particular preferred direction thereby encoding spatial information into a phase code (see Figure 20).

One advantage of describing the depolarizing effect of VCOs in an abstract manner is the simple interpretation of the model parameters. For instance, the spacing of the grid can be read out from the model by computing the distance traversed between maxima of the interference pattern. The distance in any straight direction is $L(\phi - \phi_d) = s(t)/|f_a(t) - f_b(t)| = 1/\beta|\cos(\phi - \phi_d)|$, which gives the minimum distance $L = 1/\beta$ as the spacing between parallel stripes. The grid spacing is the distance between peaks of the intersection of several VCO stripe patterns, and would thus be given by $G = 2/\sqrt{3}\beta$.

PART II

RESULTS

In the following chapters the main results of this doctoral work are presented in the form of academic research manuscripts. Preceding each of the manuscripts the reader can find a brief overview summarising the main findings to be discussed in detail in the pages to come. The purpose of these overviews is to offer the reader in clear and simple terms what are the motivations behind each of the works. They will be introduced to a general statement of the problems to be solved, why are they relevant and of interest to the neuroscience community, and to what kind of questions the reader is expected to find some answers hiding within the details.

In the first chapter we face the problem of how best to integrate information coming from different neural pathways. Diverse types of information usually have different biological requirements for encoding. We show how the biological set-up of interacting networks reflect these needs.

In the second chapter we focus on the problem of representation of space in the hippocampus. A spatially broad place cell code begs the question whether any specific location is given a special role. This question and its possible answer are contrasted with experimental findings.

The third chapter explores the reasons behind the appearance of hexagonality in the firing patterns of grid cells. Its origin is traced back to a special relationship in the spiking activity of place cells. We show how the principles leading to the formation of these characteristic patterns are commonplace in nature.

On spatial memory

It is not difficult to understand why you might feel overwhelmed when walking into a new place for the first time. Depending on the situation you find yourself in, it might be of the most urgent importance to have in your mind a quick and reliable map of your surroundings, in some cases a matter of survival.

But knowing what is where is not a simple task for the brain to accomplish. In particular, there is always a trade-off between a neural representation that is both quick to build and at he same time reliable, in other words encoding the highest amount of relevant information in a limited time. The trade-off is evident when the changes required to build the new representation come at high expenditure of biological and metabolic resources, as could be the case when the changes come in the form of synaptic plasticity in the brain.

Fortunately all representations need not be equally expensive. Every time an object is moved around there is no need to forget its previous existence and create a new memory of the object and its new location. The continuity of an object's identity throughout a set of experiences suggests that it is best to encode apart its identity and location. But is the brain making use of such encoding scheme?

Hippocampal area CA1 has been shown to encode both an object's location and identity. They are provided via the convergence of two main neuronal input streams, area CA3 and the LEC. Area CA3 sparsely encodes information about an object's location, further transmitted to CA1 through highly plastic projections. By contrast, the LEC tends to encode objects identities densely, and inducing plastic changes in its connections to CA1 may require engaging extra energy expenditure triggered by novelty (new objects).

Is this biological set-up consistent with the idea that for such an encoding task the most optimal arrangement is also the one requiring the less energy consumption?. In the following letter we show that this is indeed the case, and moreover we generalize our results to include the description of other brain areas receiving multiple incoming pathways.

All contributions to the following work were done under the supervision of Christian Leibold. C.L. and M.M. formulated and developed the analytical model. C.L. performed the numerical analysis. All authors wrote and reviewed the manuscript.

The results were presented by C.L. as a short talk at the Bernstein Conference 2015 in Heidelberg.

The results were published in the journal Neural Computation with the following reference

C. Leibold and M. M. Monsalve-Mercado. 2016. "Asymmetry of neuronal combinatorial codes arises from minimizing synaptic weight change". *Neural Computation* 28 (8). ISSN: 1530888X. doi:10.1162/NECO{_}a{_}00854

LETTER Communicated by Yasser Roudi

Asymmetry of Neuronal Combinatorial Codes Arises from Minimizing Synaptic Weight Change

Christian Leibold
leibold@bio.lmu.de
Department Biology II, Ludwig-Maximilians-Universität München, and Bernstein Center for Computational Neuroscience Munich, 82152 Martisreid, Germany

Mauro M. Monsalve-Mercado
monsalve@bio.lmu.de
Department II Biology, Ludwig-Maximilians-Universität München, 82152 Martisreid, Germany

Synaptic change is a costly resource, particularly for brain structures that have a high demand of synaptic plasticity. For example, building memories of object positions requires efficient use of plasticity resources since objects can easily change their location in space and yet we can memorize object locations. But how should a neural circuit ideally be set up to integrate two input streams (object location and identity) in case the overall synaptic changes should be minimized during ongoing learning? This letter provides a theoretical framework on how the two input pathways should ideally be specified. Generally the model predicts that the information-rich pathway should be plastic and encoded sparsely, whereas the pathway conveying less information should be encoded densely and undergo learning only if a neuronal representation of a novel object has to be established. As an example, we consider hippocampal area CA1, which combines place and object information. The model thereby provides a normative account of hippocampal rate remapping, that is, modulations of place field activity by changes of local cues. It may as well be applicable to other brain areas (such as neocortical layer V) that learn combinatorial codes from multiple input streams.

1 Introduction

With their dendrites, neurons sample from multiple anatomically and functionally distinct input pathways and process this information in spatially and computationally segregated compartments (Schaefer, Larkum, Sakmann, & Roth, 2003; Spruston, 2008; Torben-Nielsen & Stiefel, 2010; Medan & Preuss, 2014). These distinct dendritic compartments receive inputs from populations with different firing statistics and also can show different forms of synaptic plasticity (Letzkus, Kampa, & Stuart, 2006; Johenning et al., 2009;

Neural Computation **28, 1527–1552** (2016)
doi:10.1162/NECO_a_00854

Ito & Schuman, 2012). So far it is unclear whether there are generic design principles underlying the spatial segregation of different forms of synaptic plasticity and how those might relate to the neuronal code. In this letter, we propose an analytically tractable mathematical model that predicts an optimal network design and coding principles for a neuronal population that combines two input streams. In particular, the model shows that a separation into a less plastic and a more plastic pathway is beneficial if the system is to minimize the total cost of synaptic change. This objective is supposed to be essential if ongoing synaptic plasticity has to reflect memory formation of ongoing changes of the external world. The model predicts that the more plastic pathway should be encoded more sparsely than the less plastic one.

As an illustrative example, we consider hippocampal distal CA1 pyramidal cells receiving inputs from layer III of the lateral entorhinal cortex (LEC), in contrast to proximal CA1 neurons that receive inputs from layer III of the medial entorhinal cortex (MEC) (Moser et al., 2014). Layer III LEC inputs arrive at the distal apical dendrites in stratum lacunosum moleculare (lm) and are believed to convey information on local objects (Neunuebel, Yoganarasimha, Rao, & Knierim, 2013; Deshmukh & Knierim, 2011). In addition CA1 pyramidal cells receive a prominent input at the proximal apical dendrite in stratum radiatum from CA3 via the highly plastic Schaffer collateral (SC) synapses conveying mostly spatial information. The distally arriving lm inputs are also plastic; however, their synapses change less readily and may require additional release of neuromodulators, which has led to the hypothesis that these synapses change only if a novel and important input pattern arrives via this pathway (Golding, Staff, & Spruston, 2002; Sajikumar & Korte, 2011; Ito & Schuman, 2012). Since particularly small objects frequently change their position in space, forming memories of object locations should require the least possible amount of synaptic resources, and thus our theory will address why particularly the spatial (SC) input pathway is more plastic than the object (lm) pathway.

A further puzzling problem concerning the CA1 principal neurons arose with the accounts of spatial activity patterns in the MEC (Fyhn, Molden, Witter, Moser, & Moser, 2004). Particularly MEC grid cells provide a most efficient neuronal space representation (Mathis, Herz, & Stemmler, 2012) and therefore the reason for having the much sparser hippocampal place code in addition to the efficient grid code became unclear (Kammerer & Leibold, 2014). One hypothesis to resolve this apparent duplicity was that grid cells provide a genuine space representation, whereas hippocampal place cells connect space and object information (Hartley, Lever, Burgess, & O'Keefe, 2014; Neunuebel et al., 2013). In short, this idea can be paraphrased as "CA1 encodes what is where." Although this is somewhat simplified (Knierim, Neunuebel, & Deshmukh, 2014), we use this paraphrase as a framework for a thought experiment to understand how a neural network should most efficiently combine two distinct sources of input that reflect

a highly dynamic outside world with objects continuously changing their position in time. We show that hippocampal distal CA1 pyramidal cells are in a parameter regime in which synaptic weight changes are minimized and, as a result, accounts for the observed sparseness of the CA1 place representation.

2 Results

The letter is organized as follows. First, we provide an example simulation of hippocampal rate remapping. In section 2.2, we use rate remapping to illustrate the ideas underlying a mathematical theory by which we assess the integration of two input pathways on the single neuron level. The model we use adheres to mathematical approaches to memory retrieval (Cover, 1965; Hopfield, 1982; Fusi, Drew, & Abbott, 2005; Päpper, Kempter, & Leibold, 2011; van Rossum, Shippi, & Barrett, 2012), in which the strength of the input signal that drives the recall of a memory is evaluated against the noise background arising from the interference between the memory patterns.

In section 2.3, the average weight change derived previously will be minimized under the constraint of a finite number of input synapses, and the ensuing optimal solution will be discussed regarding its neurobiological implications. Specifically, the theory predicts that asymmetric sparseness is a result of a minimization of the synaptic weight changes that are required to update the memory of object locations.

2.1 Example: Hippocampal Rate Remapping. Objects frequently change their position in space, yet we are able to form memories of where objects were located at different points in time. Neurons in the hippocampal CA1 region represent both object position by their place field firing and object identity by means of rate modulations of the place cell population (rate remapping). So far, it is unclear why one of the features (space) is encoded very sparsely while the other one (object identity) is encoded densely (e.g., Yoganarasimha, Rao, & Knierim, 2011).

In area CA1, a place cell has one to four place fields in a typical box of 1m^2 (Fyhn et al., 2004) which amounts to a sparseness of about $f_1 = 10^{-1}$ (estimated for CA3 in Kammerer & Leibold, 2014), indicating that at each spatial position, the fraction 0.1 of all CA1 neurons is active. Place fields most likely reflect the excitatory input at the proximal apical dendrite of the CA1 pyramidal cell (stratum radiatum). The distal apical input (stratum lacunosum moleculare) is generally assumed to trigger a phenomenon called rate remapping (Leutgeb, Leutgeb, Treves, Moser, & Moser, 2004) in which place fields remain stable but change their firing rate under local changes of the environment. Typically such a local change is induced by altering the color of (parts of) the box in which the experiment takes place while repeating the experiment in the same laboratory room. More substantial

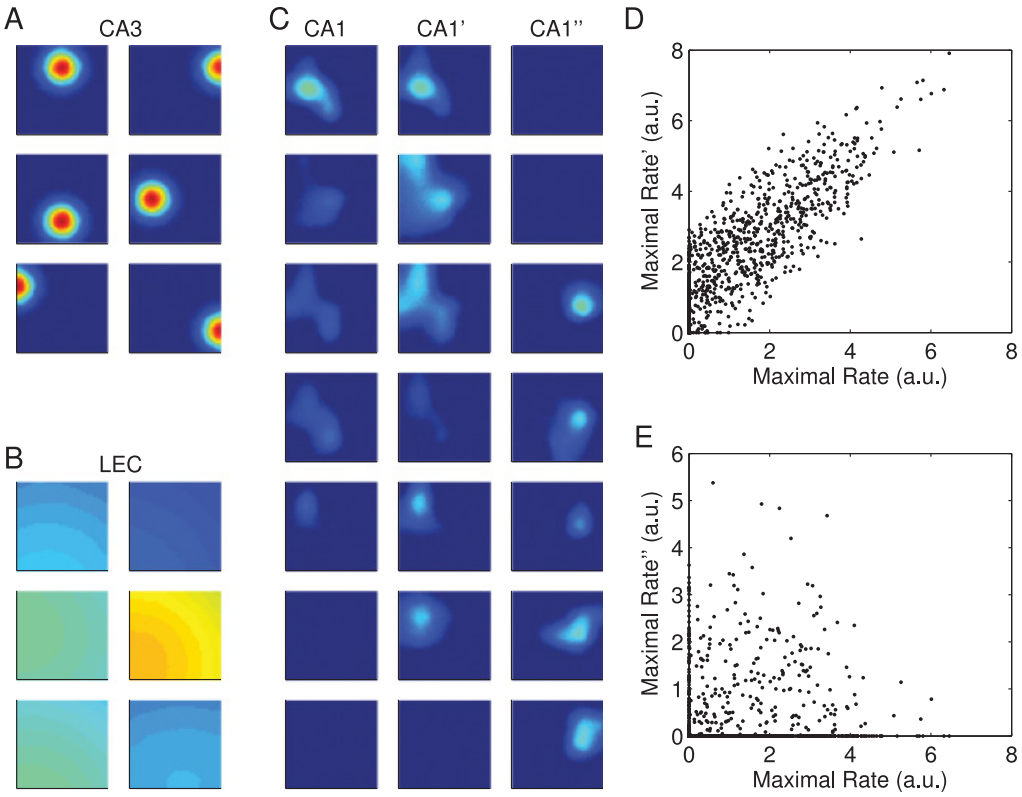


Figure 1: Hippocampal remapping. (A) Place cells (six examples) provide spatial input. (B) Object encoding cells (six examples) are only little spatially modulated. (C) CA1 cells (seven examples) in a first environment 1 (left), in the same environment with one object exchanged by another object (middle), and in a different second environment. (D) Rate remapping: Maximal firing rates of all CA1 cells in environment 1 versus environment 1 with one object exchanged. (E) Global remapping maximal firing rates of all CA1 cells in environment 1 versus environment 2.

changes of the environment (such as repeating the experiment in a different room) lead to changes in the place field positions (relative to the box), a phenomenon termed global remapping (Leutgeb et al., 2004).

The sparsely encoded spatial inputs arrive via the plastic Schaffer collateral synapses, whereas the dense object-encoding inputs arrive via the less plastic temporoammonic synapse. So far it is unclear what the functional purpose of the distinct codes and plasticity properties is. We propose that the network architecture observed in hippocampal CA1 that underlies rate remapping results from an optimality principle outlined in the next sections.

The simulations from Figure 1 are performed according to appendix B. Examples of space- and object-encoding firing fields in the input populations are shown in Figures 1A and 1B, respectively. These two inputs

generate putative CA1 place fields (see Figure 1C) that undergo rate remapping (see Figure 1D) if only one of five objects is exchanged in a known environment. Since every object drives a large fraction (in this case 1/2) of the neurons, the other four unchanged objects account for largely stable response profiles, whereas the exchange of one object can be observed only via rate modulations. Conversely, if a completely different environment is simulated (different objects and different place code), the code undergoes global remapping (see Figure 1E).

From the example of the hippocampal object and place code, we derive several important conclusions and constraints for our model:

- Object identity and position are independent variables. Objects may appear, disappear, and be dislocated in space. A learning rule for object position should capture these three situations.
- For building a new memory of an object position, neither the object nor the spatial environment has to be novel; it is sufficient to move a known object in a known environment. We thus have to distinguish this task from learning that occurs when a novel object or a novel spatial enclosure is first encountered.
- Since objects may change their position in space frequently, we assume that the amount of synaptic resources invested in learning such changes should be as small as possible, such that most plasticity resources can be used for learning novel items and places, which occurs only once per such an event.
- The problem can be mapped to a combinatorial code. If there exist neuronal representations (patterns) for P_1 positions and P_2 objects, the input space to be considered consists of the product of $P_1\,P_2$ possible object locations. The numbers P_i should thus serve as parameters that define the complexity of the task.
- The sparseness f_i of the input pathways quantifies the fraction of patterns a neuron is excited by; that is, $Q_i = f_i\,P_i$ defines the number of such patterns. In the case of spatial position, Q and f are small (sparse code); for object identity, Q and f are large (dense code). From binomial statistics, the expected number of active inputs per pathway then equals $f_i\,N_i$, and thus f_i also describes the average fraction of active neurons in the population.
- The neuronal representation of object locations must occur at the level of the population of neurons that combines the two input streams (in the example of the hippocampus CA1). A particular identified activity pattern in this readout population encodes for a learned combination of object place and identity. If an object does not occur at the stored location, no such identified pattern will be active.

2.2 Model Design. If a CA1 pyramidal cell encodes for a combination of place and objects, it should be active only if both stimuli, the position

stimulus and the right object, are present, whereas it should remain silent if only the place stimulus is correct or only the object is correct. From the perspective of the spike generator, this means that both inputs individually should remain subthreshold, whereas the sum of the two inputs is suprathreshold. Mathematically the depolarization v of the CA1 neuron is modeled as a linear superposition of both input streams,

$$v = \vec{w}_1 \cdot \vec{x}_1 + \vec{w}_2 \cdot \vec{x}_2, \tag{2.1}$$

where the N_i-dimensional vector $\vec{x}_i$ denotes the firing rates of the N_i input neurons from population i, and $\vec{w}_i$ denotes the respective synaptic weights. Which of the two input pathways corresponds to space and which one to objects is not further specified in the model design. As a result of the optimization process described in the following sections, it will turn out that pathway 1 describes the synapses conveying spatial information, and pathway 2 labels the synapses conveying object information.

The pathway-specific synaptic weights are chosen according to a Hebbian rule as

$$\vec{w}_1 = \sum_{n=1}^{Q_1} \vec{x}_1^{(n)}, \quad \vec{w}_2 = \sum_{m=1}^{Q_2} \vec{x}_2^{(m)}. \tag{2.2}$$

The weight vectors from equation 2.2 are derived from the two sets of patterns $\{\vec{x}_1^{(1)}, \ldots, \vec{x}_1^{(Q_1)}\}$ and $\{\vec{x}_2^{(1)}, \ldots, \vec{x}_2^{(Q_2)}\}$, to which a neuron should elicit a depolarizing response of its membrane potential. These Q_i ON patterns are only a fraction $f_i = Q_i/P_i$ of the total number P_i of patterns transmitted by the individual pathways. The remaining $(1 - f_i)\, P_i$ OFF patterns that the neuron should not respond to are supposed to be ON patterns for other neurons. Thus, a large population of output neurons that randomly sample ON and OFF patterns is able to convey the information about which of the $P_1\, P_2$ combinations is present in the input—for the CA1 example, the locations of the available objects.

The fractions $f_i = Q_i/P_i$ of ON patterns are called pathway-specific sparseness values, since low f_i will translate to an output population with only a few neurons active.

The difference between the depolarization resulting from two correct inputs and the depolarization resulting from only one correct input can be considered to be the signal that drives the neuron to fire. Assuming a fixed probability of an erroneous response requires a constant signal-to-noise ratio $\sqrt{K}$ (see Figure 2), that is, the noise has to be proportional to the signal according to

$$\text{Signal} = \sqrt{K}\,\text{Noise}. \tag{2.3}$$

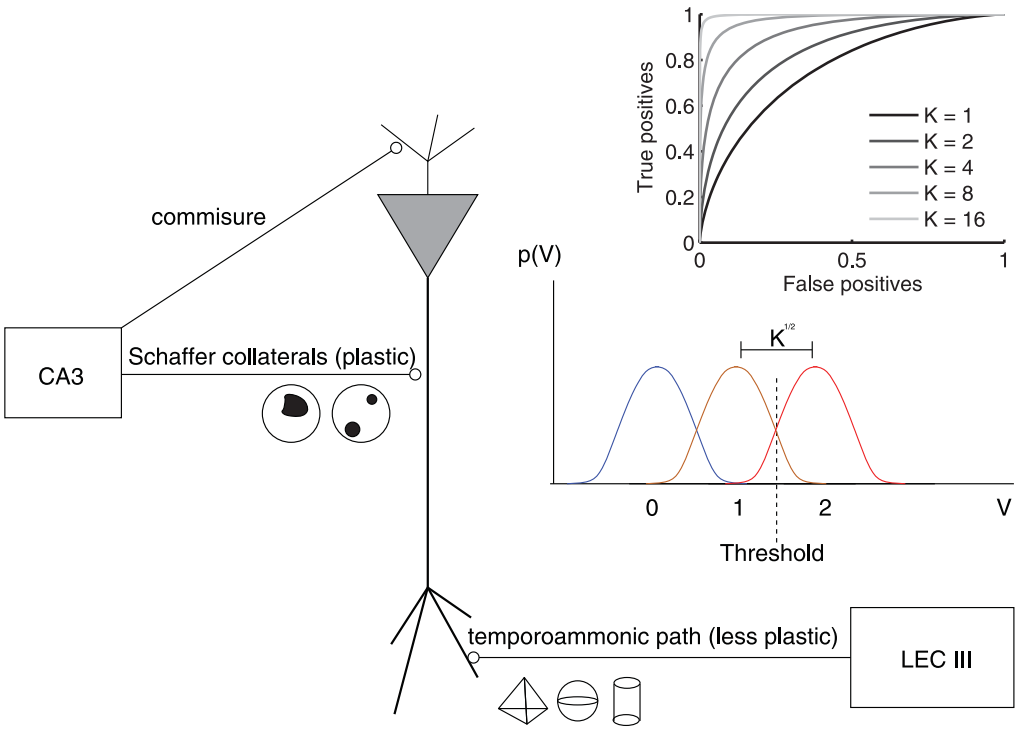


Figure 2: Anatomical separation of input pathways. The long apical dendrite of a hippocampal pyramidal cell allows for a spatial separation of inputs from different areas. For distal (close to subiculum) CA1 neurons, the inputs in stratum radiatum (sr) convey spatial information from ipsilateral CA3 (contralateral CA3 projects to basal dendrites via the commisure). The inputs at the distal apical dendrite in stratum lacunosum moleculare (lm) arise from layer III of the lateral entorhinal cortex and convey object information. Proximal sr synapses are highly plastic, whereas plasticity at lm synapses requires conjunctive modulatory action. The postsynaptic voltage response V results from both input pathways. We assume that the neuron is most depolarized if both pathways are appropriately activated (red voltage distribution). Activation of only one (brown distribution) or no (blue distribution) pathway leads to lower depolarization. The firing threshold (dashed line) is thus assumed to optimally segregate the voltage distributions for one and two active inputs. The error probabilities of the suprathreshold response are determined by the (noise normalized) separation $\sqrt{K}$ of the two distributions, as indicated by the ROC curves.

For large K, the depolarization difference between one and two correct inputs is large and the number of erroneous responses will be small, and vice versa. For large numbers of input synapses, the membrane depolarizations will approximately follow a gaussian distribution according to the central limit theorem. Then the probabilities that the output neuron responds wrongly can be computed from standard signal detection theory (Peterson,

Birdsall, & Fox, 1954). Given a firing threshold Θ and the signal-to-noise ratio $\sqrt{K}$, the probability of a missed spike (false negative) equals

$$p_{\text{miss}} = \frac{1}{2}\left[1 + \text{erf}\left(\frac{\Theta - \sqrt{K}}{\sqrt{2}}\right)\right], \tag{2.4}$$

whereas the probability for a wrong spike (false positive) equals

$$p_{\text{err}} = \frac{1}{2}\left[1 - \text{erf}\left(\frac{\Theta}{\sqrt{2}}\right)\right]. \tag{2.5}$$

By varying the firing threshold Θ, the two equations yield the usual ROC curves (see Figure 2, top), which have a steeper increase for larger signal-to-noise ratio $\sqrt{K}$.

2.2.1 Signal. Without loss of generality, we use equation 2.1 to compute the depolarization $v^{(1,1)}$ resulting from two ON input patterns as the response to the first two input patterns the neurons has learned to respond to, that is, $\vec{x}_1 = \vec{x}_1^{(1)}, \vec{x}_2 = \vec{x}_2^{(1)}$,

$$v^{(1,1)} = |\vec{x}_1^{(1)}|^2 + \sum_{n=2}^{Q_1} \vec{x}_1^{(n)} \cdot \vec{x}_1^{(1)} + |\vec{x}_2^{(1)}|^2 + \sum_{m=2}^{Q_2} \vec{x}_2^{(m)} \cdot \vec{x}_2^{(1)}. \tag{2.6}$$

The average suprathreshold response (expected value with respect to the probability distributions $p_i(x_i)$ of input rates x_i in pathway i) therefore amounts to

$$\langle v^{(1,1)} \rangle = N_1 \,\text{var}(x_1) + Q_1 N_1 \langle x_1 \rangle^2 + N_2 \,\text{var}(x_2) + Q_2 N_2 \langle x_2 \rangle^2. \tag{2.7}$$

Note that for reasons of simple notation, $\text{var}(x_1)$ denotes the variance with respect to the distribution $p_1(x_1)$ of firing rates in pathway 1. Similarly, computing the average depolarization $v^{(0,0)}$ to a combination of patterns $\vec{x}_1^{(0)}$ and $\vec{x}_2^{(0)}$ that are not contributing to either synaptic weight vector can be obtained as

$$\langle v^{(0,0)} \rangle = Q_1 N_1 \langle x_1 \rangle^2 + Q_2 N_2 \langle x_2 \rangle^2. \tag{2.8}$$

Hence, the difference between the suprathreshold response and this subthreshold response,

$$\langle v^{(1,1)} - v^{(0,0)} \rangle = N_1 \,\text{var}(x_1) + N_2 \,\text{var}(x_2), \tag{2.9}$$

is a sum of two signals,

$$\text{Signal}_i := N_i \,\text{var}(x_i),\, i = 1, 2,$$

each of which relates to one of the input pathways. Analogously, the depolarizations $v^{(1,0)}$ and $v^{(0,1)}$ resulting from an ON input in one pathway and an OFF input in the other pathway are

$$\langle v^{(1,0)} \rangle = \langle v^{(0,0)} \rangle + \text{Signal}_1,\ \langle v^{(0,1)} \rangle = \langle v^{(0,0)} \rangle + \text{Signal}_2. \tag{2.10}$$

Since space is continuous, the spatial input patterns will not always perfectly be present; rather, they gradually decline as the animal moves away from an ON position. If we parameterize this gradual decline by η (where $\eta = 0$ denotes the perfect pattern and $\eta = 1$ indicates a decorrelated pattern), analogous mathematical steps reveal that the signal rescales like $\text{Signal} \to \text{Signal}\,(1 - \eta)$. Under the same assumptions that lead to equations 2.4 and 2.5, the probability that a neuron fails to fire thus increases with η like

$$p_{\text{miss}}(\eta) = \frac{1}{2}\left[1 + \text{erf}\left(\frac{\Theta - (1-\eta)\sqrt{K}}{\sqrt{2}}\right)\right]. \tag{2.11}$$

Since the error function in equation 2.11 is continuous in η, the response rate of a place cell will thus smoothly decline as the animal moves away from the place field center. In what follows, we therefore consider space to be represented by uncorrelated input patterns encoding discrete positions.

2.2.2 Noise. The variance (noise) of the depolarization,

$$\begin{aligned}\langle (v^{(0,0)} - \langle v^{(0,0)} \rangle)^2 \rangle = {} & N_1\, Q_1\, \text{var}(x_1)[\text{var}(x_1) + \langle x_1 \rangle^2 (Q_1 + 1)] \\ & + N_2\, Q_2\, \text{var}(x_2)[\text{var}(x_2) + \langle x_2 \rangle^2 (Q_2 + 1)],\end{aligned} \tag{2.12}$$

is minimal for zero-mean distributions $p_i(x_i)$: $\langle x_1 \rangle = \langle x_2 \rangle = 0$. Further on, it is assumed that these two expected values vanish by a pathway-specific balance between excitation and inhibition (Monier, Fournier, & Fregnac, 2008). Moreover, the total Noise is assumed to be composed of the minimal variance from equation 2.12 and an internal voltage jitter of variance σ^2:

$$\text{Noise} = \sqrt{N_1\, Q_1\, \text{var}(x_1)^2 + N_2\, Q_2\, \text{var}(x_2)^2 + \sigma^2}. \tag{2.13}$$

2.2.3 Minimal Signal. The decision of the neuron to fire an action potential or not is a decision between one or two ON patterns as inputs. Assuming

that the two Signals in the two pathways are of different size and that the signal-to-noise ratio $\sqrt{K}$ should be fixed, the Noise from equation 2.13 must be low enough to fit the signal-to-noise condition for the pathway with the smaller Signal. If without loss of generality $\text{Signal}_1 \leq \text{Signal}_2$, the condition from equation 2.3,

$$\text{Signal}_1 = \sqrt{K}\,\text{Noise}, \tag{2.14}$$

then means that the minimal "Signal" (Signal_1) on which a neuron has to decide whether to respond corresponds to detecting whether pathway 1 is correctly activated in the case that pathway 2 is correctly activated.

Introducing the abbreviation $q_i = Q_i/N_i$, equations 2.13 and 2.14 read

$$(\text{Signal}_1)^2 = K\,(q_1\,(\text{Signal}_1)^2 + q_2\,(\text{Signal}_2)^2 + \sigma^2) \tag{2.15}$$

or

$$\text{Signal}_1 = \sqrt{\frac{K\,q_2\,(\text{Signal}_2)^2 + K\,\sigma^2}{1 - K\,q_1}}. \tag{2.16}$$

The term under the square root must not be negative and thus

$$q_1 < 1/K.$$

Since $(\text{Signal}_1)^2 \leq (\text{Signal}_2)^2$, we obtain

$$(\text{Signal}_2)^2 \geq \frac{K\,\sigma^2}{1 - K\,(q_1 + q_2)} \tag{2.17}$$

and the further constraints

$$q_1 + q_2 < 1/K \;\Rightarrow q_2 < 1/K - q_1 \Rightarrow q_2 < 1/K.$$

From equations 2.16 and 2.17 it follows that $(\text{Signal}_1)^2$ reaches its infimum $\frac{K\,\sigma^2}{1-K\,(q_1+q_2)}$ if both signals, Signal_1 and Signal_2, are equal. The minimal Signal is thus given by

$$\text{Signal} = N_1\,\text{var}(x_1) = N_2\,\text{var}(x_2) = \sqrt{\frac{K\sigma^2}{1 - K\,(Q_1/N_1 + Q_2/N_2)}}. \tag{2.18}$$

In what follows, we will use equation 2.18 to compute the variance of the inputs $\text{var}(x_i)$, and thus the neuron operates in a regime in which the

decision of firing is based on a minimal signal. The minimum of the average synaptic weight change, which we assume is the primary objective of the learning rule, will be shown to occur for this minimal Signal.

A direct conclusion from equation 2.18 is that $Q_1/N_1 + Q_2/N_2 = 1/K$ defines a capacity supremum for the numbers Q_1 and Q_2 of ON patterns. At $Q_1/N_1 + Q_2/N_2 = 1/K$ interference induces such large Noise that the Signal needs to become infinite to still be able to account for the finite error rate imposed by K. Moreover, the numbers $Q_{1,2}$ of patterns the neuron responds to scale with the respective number of input dimensions $N_{1,2}$ (i.e., a linear increase in $N_{1,2}$) is sufficient to achieve larger values $Q_{1,2}$ while keeping the Signal the same.

2.2.4 What Are the Independent Parameters?. To discuss further implications of equation 2.18, we need to identify which parameters are free (independent) and which ones can be obtained by optimality principles. We consider the numbers $N_{1,2}$ of input synapses, the parameters P_i describing the respective input pathways, and the parameters constraining the code of the output population:

- *Number of synapses.* While in most models, the network size is a constant parameter that sometimes is fitted to the assumed biological reality, the framework presented here starts with no a priori assumption about the number of input neurons. This is particularly important since the numbers N_1 and N_2 of input neurons are expected to be strongly constrained by the requirements of the input codes and the learning rule. Both N_1 and N_2, will be obtained by concurrently minimizing the total input dimension $N = N_1 + N_2$ and the amount of synaptic change.
- *Input code.* In this letter the information content of the input pathways is quantified by the numbers P_1 and P_2 of distinct input patterns $\vec{x}_1$ and $\vec{x}_2$ these pathways can provide. The product $P_1\,P_2$ thus amounts to the total number of combinations the neuron can be confronted with. In the "what is where" example, $P_1\,P_2$ is the number of possibilities to distribute a set of P_2 known objects in a familiar space or, more precisely, a discrete set of P_1 known spatial positions. The numbers P_1 and P_2 are two free parameters of the model describing the behavioral demands implicated by the world the hippocampus encodes.
- *Output code.* The signal-to-noise ratio $\sqrt{K}$ has already been introduced to parameterize the code in the output (CA1) population. It determines the fraction of wrong decisions (fire or not fire). The combinatorial code, however, also requires specifying how often a neuron should receive a depolarizing signal, that is, the fractions $f_{1,2}$ of the $P_{1,2}$ stimuli the synapses have been trained to listen to. Assuming that the synaptic changes in the two pathways are statistically independent, the overall sparseness of the CA1 code equals

$$f = f_1 f_2. \tag{2.19}$$

The readout also imposes a third parameter, which is the upper bound on sparseness $\varphi \geq f_i$, since the overall information content of the population code drops to zero if f_i converges to 1 or 0. Since $f_{1,2}$ are related via $f = f_1 f_2$, the upper bound φ also induces a lower bound $f/\varphi \leq f_i$. Since no further assumptions will be made on the readout of the output population, the three parameters K, f, and φ of the output code cannot be derived and will remain as free parameters of the model to be discussed quantitatively in what follows.

2.2.5 Assessing the Amount of Weight Change. If we assume that minimizing the synaptic weight change per memory is an important objective for strongly memory-related brain areas, we have to find an analytical expression for the expected weight change during learning a new memory.

In our framework, learning means that a neuron may either no longer be depolarized by a pattern (the pattern gets deleted from the set of ON patterns), be depolarized by a pattern that was previously an OFF pattern (insertion of an ON pattern), or a previous ON pattern might be exchanged by a new ON pattern (replacement). In the example of memorizing object location, it would mean that when an object gets relocated (and a memory of the new object position has to be formed), the input pattern corresponding to the old position has to be deleted from the weight vector of the neurons that encode object identity and the old position. Similarly, the input pattern corresponding to the new position has to be inserted to the weight vector of the neurons that encode object identity and the new position. Finally, the input pattern of the old position has to be replaced by the input pattern of the new position for neurons that encode object identity and both positions.

Assuming that each neuron is randomly assigned Q_i ON patterns from the fixed set of P_i patterns, the probability that a neuron responds to a specific input pattern is $f_i = Q_i/P_i$. Thus, deletion (D), insertion (I), or replacement (X) of an ON pattern occurs at a single neuron with probability $f_i\,(1 - f_i)$ for deletion and insertion ($1 - f_i$ is the probability that one pattern, new or old, was an OFF pattern, and f_i is the probability that the other pattern is an ON pattern) and with probability $(f_i)^2$ for replacement (both patterns, new and old, are ON patterns).

If, during learning, an input pattern $\vec{x}_i$ is exchanged (X) by another pattern $\vec{x}_i'$, the overall synaptic change equals,

$$\Delta^{\mathrm{X}} w_i = |\vec{x}_i - \vec{x}_i'| \Rightarrow \langle (\Delta^{\mathrm{X}} w_i)^2 \rangle = 2\, N_i\, \mathrm{var}(x_i),$$

where $\langle \cdot \rangle$ and var denote the expected value and variance over the distribution $p_i(x_i)$ of input rates (vector elements of $\vec{x}_i$) in pathway i. Exchanging an input pattern is assumed to occur with probability $(f_i)^2$.

If a new input pattern is introduced (I) or an existing input pattern is deleted (D), the corresponding change in the weight vector equals

$$\Delta^{\mathrm{I,D}} w_i = |\vec{x}_i| \Rightarrow \langle (\Delta^{\mathrm{I,D}} w_i)^2 \rangle = N_i \, (\mathrm{var}(x_i) + \langle x_i \rangle^2).$$

Considering all three types of weight change (D,I,X) and their respective probabilities, we obtain the average weight change as

$$\langle \Delta w_i \rangle = f_i \, (1 - f_i) \, \langle \Delta^{\mathrm{I}} w_i \rangle + f_i \, (1 - f_i) \, \langle \Delta^{\mathrm{D}} w_i \rangle + (f_i)^2 \, \langle \Delta^{\mathrm{X}} w_i \rangle. \quad (2.20)$$

Owing to Jensen's inequality, ($\langle \sqrt{a^2} \rangle \leq \sqrt{\langle a^2 \rangle}$), the average synaptic change is bounded by

$$\begin{aligned} \langle \Delta w_i \rangle &\leq 2 \, f_i \, (1 - f_i) \sqrt{(\langle \Delta^{\mathrm{I,D}} w_i)^2 \rangle} + (f_i)^2 \sqrt{(\langle \Delta^{\mathrm{X}} w_i)^2 \rangle} \\ &= 2 \, f_i \sqrt{N_i \mathrm{var}(x_i)} \left[(1 - f_i) \sqrt{1 + \frac{\langle x_i \rangle^2}{\mathrm{var}(x_i)}} + f_i / \sqrt{2} \right]. \end{aligned} \quad (2.21)$$

Since the number of input synapses will be large, the relative variance of $\Delta^{(j)} w_i$ will be small and thus the upper bound will be a good approximation to $\langle \Delta w_i \rangle$.

From equation 2.21 and the assumption of balanced inputs $\langle x_i \rangle = 0$, we find that the average synaptic change in pathway i equals

$$\langle \Delta w_i \rangle = 2 \sqrt{\mathrm{Signal}_i} \, f_i \, (1 - f_i (1 - 1/\sqrt{2})) = 2 \sqrt{\mathrm{Signal}_i} \, f_i \, (1 - \gamma \, f_i),$$

where $\gamma \approx 0.3$. If f_i is small, we may further on neglect the last factor and use

$$\langle \Delta w_i \rangle^4 \approx 16 \, (\mathrm{Signal}_i)^2 \, (f_i)^4 = \frac{16 \, K \sigma^2 \, (f_i)^4}{1 - \kappa_1 f_1 - \kappa_2 \, f_2} \quad (2.22)$$

as the objective function to minimize synaptic change, with $\kappa_i = K P_i / N_i$. For reasons of analytical tractability, we will consider the fourth power of $\langle \Delta w_i \rangle$, since applying a monotonous function does not change the location of the minimum and, in case of constraints, only rescales Lagrange multipliers.

2.2.6 Result 1: Sparse Pathway Is More Plastic. Equation 2.22 can be intuitively understood since, first, only the fraction f_i of neurons that receive an ON input need to be affected by synaptic learning. If the neurons in the output population randomly sample their ON input patterns, the mean fraction of neurons receiving an ON input will match the fraction $f_i = Q_i / P_i$ of ON patterns due to binomial statistics. Second, the synaptic change in

the learning rule from equation 2.2 equals the input rate vector that also constitutes the strength of the Signal.

Minimizing weight change thus means both Signal and sparseness f_i have to be minimized. Minimizing the Signal yields a pathway-independent value (see equation 2.18). Thus, under this condition, the distinction of the amounts of synaptic change required in each individual pathway is made only by the pathway-specific sparseness f_i. Since only one pathway has to be plastic to be able to learn all of the $P_1 P_2$ combinations of patterns, equation 2.18 has the important implication that the minimization of synaptic changes predicts that plasticity must occur at the pathway with the lowest sparseness f_i. Without loss of generality, we further on assume $f_1 < f_2$, and hence define pathway 1 as being plastic and pathway 2 as being the nonplastic one.

2.3 Optimal Networks

2.3.1 Minimizing Weight Changes. Combining equations 2.18, 2.22, and $f_2 = f/f_1$ yields (see appendix A) an explicit expression for the average change of synaptic weights,

$$\langle \Delta w \rangle = \mathcal{W}(f_1, N_1, N_2; f, K, P_1, P_2), \tag{2.23}$$

as a function of the four parameters f, K, P_1, P_2 describing output and input constraints and the three parameters f_1, N_1, N_2 that shall be found by minimization of the expected weight change $\langle \Delta w \rangle$. Interestingly, all pattern numbers P_i in equation 2.23 are scaled with K/N_i. This scaling allows the introduction of auxiliary variables $\kappa_i = K P_i/N_i$ that can be interpreted as the information per synapse in pathway i. The (fourth power of the) weight change therefore depends on only one free parameter f,

$$\langle \Delta w \rangle^4 = W(f_1, \kappa_1, \kappa_2; f) = \frac{16 K \sigma^2 (f_1)^4}{1 - \kappa_1 f_1 - \kappa_2 f/f_1}, \tag{2.24}$$

where the unknown input dimensions $N_{1,2}$ are obtained from the optimal values of $\kappa_{1,2}$.

Equation 2.24 takes a trivial minimum at $\kappa_1 = \kappa_2 = 0$, which, for finite $P_{1,2}$, corresponds to $N_{1,2} \to \infty$. To also enforce a limited total number $N_1 + N_2 = K (P_1/\kappa_1 + P_2/\kappa_2)$ of input synapses, we introduce a weight parameter λ that balances the cost of an input synapse with the cost of synaptic change by minimizing the Lagrange function,

$$\mathcal{L}(f_1, \kappa_1, \kappa_2; f, P_1, P_2, K, \lambda) = W(f_1, \kappa_1, \kappa_2; f) + \lambda K (P_1/\kappa_1 + P_2/\kappa_2). \tag{2.25}$$

The Lagrange parameter λ can thus be seen as a measure for the cost of a synapse (in units of the fourth power of synaptic change).

As discussed previously, the minimization of $\mathcal{L}$ additionally requires meeting constraints imposed by the readout, that is, the pathway-specific sparseness values have to be bounded by

$$f/\varphi \leq f_1 \leq \sqrt{f} \leq f_2 \leq \varphi. \tag{2.26}$$

The constrained optimization problem can be solved analytically for κ_1 and κ_2 and numerically or approximately analytically for f_1 and thereby yields expressions for the optimal synapse numbers N_1 and N_2, and the optimal sparseness f_1 (see below).

2.3.2 Optimal Sparseness and Input Dimensions. Minimization of the objective functions (see appendix A, equation A.1) reveals that the optimal synapse numbers are given by

$$N_i = K\left[Q_i + \sqrt{Q_1\,Q_2} + \sqrt{\frac{16\,\sigma^2\,(f_1)^4}{\lambda}}\,\sqrt{Q_i}\right]. \tag{2.27}$$

The optimal pathway-specific sparseness f_1 (see Figure 3A), as well as the input numbers N_1 and N_2, are increasing functions of P_2 (see Figure 3B). Moreover, the optimal configuration $(f_1)^{\text{opt}}$, N_1, and N_2 hardly depends on the cost parameter λ (see Figure 3C).

The optimal solutions can be subdivided into three regimes defined by the upper and lower bounds on the pathway-specific sparseness $f/\varphi \leq (f_1)^{\text{opt}} \leq \sqrt{f}$:

- *Sparse regime.* The regime of minimal $(f_1)^{\text{opt}} = f/\varphi$ is of particular biological interest. From equation A.7 (see Appendix A), it follows that the sparse regime is obtained for

$$\sqrt{\frac{f\,P_2}{P_1}} < f/\varphi \Leftrightarrow \frac{P_2}{P_1} < \frac{f}{\varphi^2} \Leftrightarrow Q_2 < Q_1.$$

  Since by definition $\sqrt{f} < \varphi$, the sparse regime is generally optimal if the number of patterns in the second pathway is substantially lower than in the plastic first pathway.
- *Dense regime.* Similar to the previous paragraph, equation A.7 (see appendix A) reveals that the dense regime $(f_1)^{\text{opt}} = (f_2)^{\text{opt}} = \sqrt{f}$ is optimal whenever

$$\frac{P_2}{P_1} > 1 \Leftrightarrow P_2 > P_1.$$

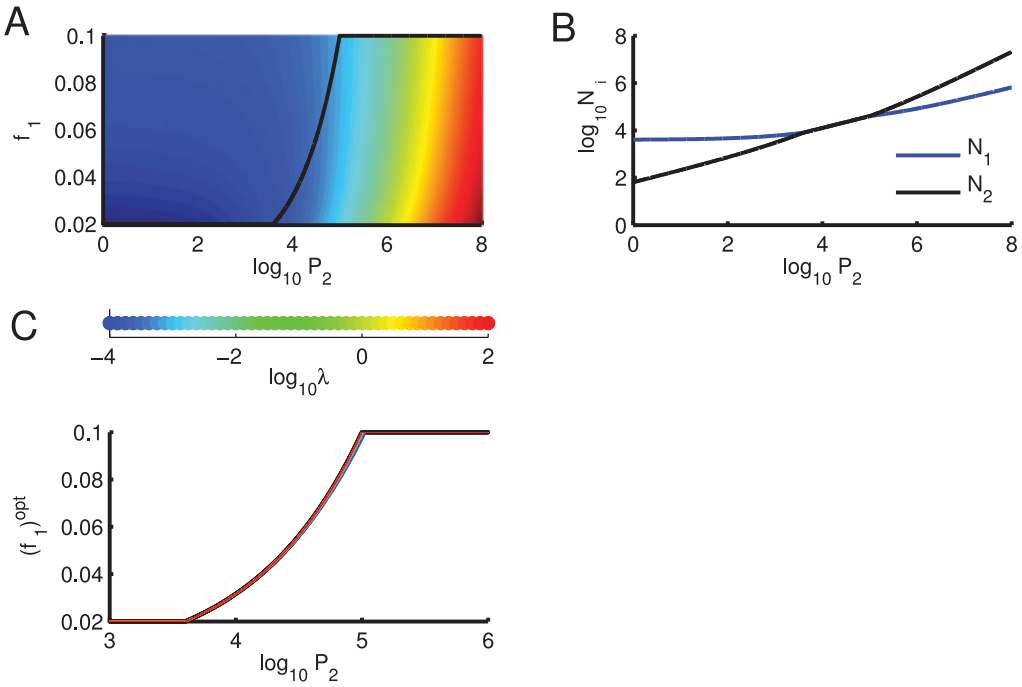


Figure 3: Optimal pathway-specific sparseness. (A) The loss function (color coded) increases with the number P_2 of inputs in the second pathway, and its minimum ($(f_1)^{\text{opt}}$ black line) with respect to f_1 is located at low values of f_1 for low P_2 and at high values of f_1 for large P_2. (B) Optimal synapse numbers $N_{1,2}$ in the two pathways as a function of P_2. Both functions are nearly identical in the intermediate regime at which $(f_1)^{\text{opt}}$ is not at the lower or upper bound; see equation 2.29. (C) Optimal pathway-specific sparseness as a function of P_2 for different synapse costs λ (in units of $(16\,\sigma^2)$, colors) and the approximation from equation 2.28, thick black line. Parameters were $P_1 = 10^5$, $\varphi = 0.5$, $f = 0.01$, $K = 2$, and $\lambda = 0.1/(16\sigma^2)$ unless otherwise mentioned.

Thus, the dense regime is optimal if the number of patterns in the nonplastic pathway is larger than that in the plastic pathway.

- *Intermediate regime.* Within the bounds $f/\varphi \le (f_1)^{\text{opt}} \le \sqrt{f}$, the optimal pathway-specific sparseness (see equation A.7) can be approximated by

$$(f_1)^{\text{opt}} = \sqrt{\frac{f\,P_2}{P_1}}. \tag{2.28}$$

The optimal input numbers N_1 and N_2 are almost identical in this regime with inactive boundary conditions (see Figure 3B) and can be analytically approximated (see equations 2.27 and 2.28) as

$$N_1 \approx N_2 \approx 2\,K\sqrt{P_1 P_2 f}. \tag{2.29}$$

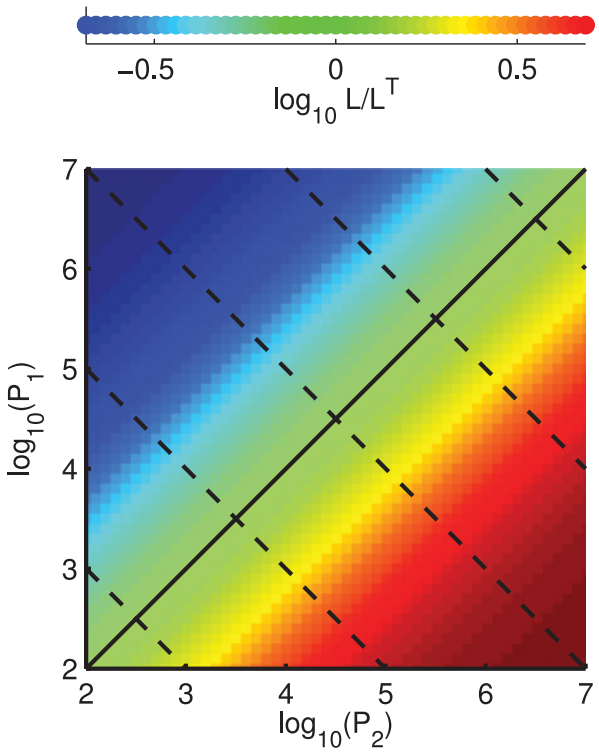


Figure 4: Effect of pattern numbers. Color encodes the ratio between optimal cost for a given P_1 and P_2 and the "transposed" cost obtained with P_1 and P_2 interchanged. The total pattern number $P_1 + P_2$ is constant along the dashed lines. The solid black line indicates identity $P_1 = P_2$ (bisecting line). The cost ratio is always smaller than 1 (logarithm smaller than 0) above the bisecting line (bluish colors). Parameters were $\varphi = 0.5$, $f = 0.01$, $K = 2$, $\lambda/(16\sigma^2) = 0.1$.

Thus, differences in P_1 and P_2 have no effect on N_i as long as their product remains constant.

2.3.3 Result 2: Learning the More Informative Pathway. So far the two input pathways are distinguished only by $f_1 < f_2$, and as a result of that, pathway 1 was concluded to be plastic and pathway 2 was defined as nonplastic (or less plastic). It is yet unclear how this plasticity property should ideally relate to the pattern numbers P_1 and P_2. The minimal cost $\mathcal{L}$ (see Figure 3A) decreases with decreasing P_2, which already indicates that the plastic pathway 1 may optimally be chosen as the one that encodes more patterns. Since, however, in Figure 3A, P_1 is kept constant, the decrease in cost may simply reflect a reduction in the overall number of inputs $P_1 + P_2$. We therefore computed the cost for all combinations of P_1 and P_2 (see Figure 4) and found that, indeed, the cost is consistently higher for $P_2 > P_1$ even if $P_1 + P_2$ is kept constant (dashed lines). This numerical observation

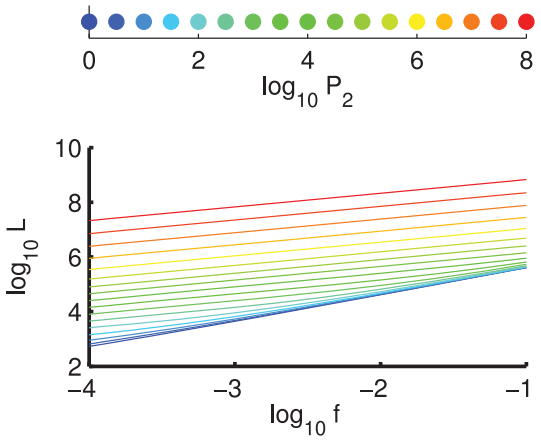


Figure 5: Sparseness helps. Minimal cost L as a function of the output sparseness f for different numbers P_2 of input patterns in the nonplastic pathway. The slopes of the curves (exponents of the power law) decrease with increasing P_2. Thus, low f is particularly cost efficient for low P_2. Parameters were $P_1 = 10^5$, $\varphi = 0.5$, $K = 2$, $\lambda = 0.1/(16\sigma^2)$.

also matches the analytical considerations from equation A.5 that show that in leading order, the cost function scales like

$$\mathcal{L} \propto \left(\sqrt{f_1 P_1} + \sqrt{f_2 P_2}\right)^2,$$

and hence the sparser pathway ($f_1 < f_2$) introduces a lower cost per pattern.

As a direct consequence, the dense regime ($P_2 > P_1$) can never be optimal, since swapping the pathways $P_1 \leftrightarrow P_2$ would reduce the loss. Thus, the pathway with lower P_i should always be plastic and the pathway with larger P_i should always be nonplastic.

The finding that the plastic (sparse) pathway should have the larger number of patterns to represent is somewhat counterintuitive since, naively, one would expect that plasticity occurring at the less informative pathway would require less effort. However, the information in pathway 2 is encoded more densely (with larger f_i) and therefore would involve a larger proportion of synapses to be changed (see equation 2.22).

2.3.4 Result 3: Sparseness Helps. So far the sparseness parameter f that has been assumed to be imposed by the readout structure was held fixed. However, the analytical results (see equation 2.27 and following) indicate that optimal networks strongly and nontrivially depend on the sparseness f. We therefore analyzed the optimal networks for varying f between 10^{-4} and 10^{-1}.

Figure 5 reveals that independent of P_2, the cost $\mathcal{L}$ increases with f and, particularly for strongly asymmetric conditions, $P_2 \ll P_1$, lowering f

(increasing sparseness) most strongly decreases the cost. Thus, the sparser the code, the fewer synapses are necessary. From the opposite perspective, the above result can also be interpreted in this way: If an observed neural code is very sparse, then the number P_1 of patterns in the plastic pathway is probably very large (and much larger than P_2).

3 Discussion

The essence of neural computation is to combine inputs from different sources to transform them into a new code that displays particular important features of the combined input space. The neuronal activity of brain areas like the hippocampus that integrate over various anatomically and functionally segregated input streams therefore must provide multimodal representations of all these input dimensions. In experimental recordings, however, only those input dimensions appear as most salient (such as space in the hippocampus) for which the neuronal output is sparse, since there, input selectivity can be easily assessed, as neurons fire for only a few stimuli and are silent otherwise. Conversely, densely represented input pathways are much harder to identify since neurons fire for a large fraction of stimuli and thus may seem to be nonselective.

In this letter, we showed by means of an analytically tractable theoretical model that if the cost of synaptic change during learning is to be minimized, a combination of a sparse and a dense input stream is most efficient for neurons combining two major input pathways. The sparser pathway thereby must be more plastic than the denser pathway. Our model suggests that only the numbers of distinct inputs determine which pathway should be plastic and predict that this should be the pathway transmitting a larger number P_i of different patterns.

We argue that distal CA1 neurons that combine space and object information are likely candidates for our model to apply. The sparse place code is conveyed via the highly plastic synapses contacting the neurons at their proximal apical dendrite. The dense object code arrives distally through less plastic synapses. The rate modulations observed during rate remapping are predicted to be a direct consequence of the dense object code since any particular neuron will be driven by a large fraction of objects and the difference between them results only in a rate modulation. If, however, a neuron is trained only to respond to a single object in a given environment, removal or replacement of this object would lead to a complete change of the place field, which is occasionally observed in experiments (Cressant, Muller, & Poucet, 1997; Leutgeb et al., 2004).

Rate remapping in the hippocampal CA1 area is much more salient than in CA3 (Lee, Yoganarasimha, Rao, & Knierim, 2004; Vazdarjanova & Guzowski, 2004; Leutgeb et al., 2004), which is generally explained by the different anatomical connection schemes in the two areas. In CA3, strong recurrent connectivity may generate salient fixed point states that

are only slightly affected by modulatory inputs but already provide a context(object)-dependent space code manner (Roudi & Treves, 2008; Solstad, Yousif, & Sejnowski, 2014). In contrast, the predominantly feedforward topology of CA1 may be more prone to rate modulations induced by the synaptic inputs at the distal dendrites and yield less stable firing fields (Mankin et al., 2012). Particularly LEC inputs to the distal CA1 region (close to the subiculum) are thought to play an important role in rate remapping (Henriksen et al., 2010), because changes of proximal cues (objects) are the standard paradigm to induce rate remapping and object-selective cells are a hallmark of LEC (Yoganarasimha et al., 2011; Deshmukh & Knierim, 2011; Lu et al., 2013).

In contrast to the distal CA1 area, the proximal CA1 area (close to CA2) receives distal inputs from the medial entorhinal cortex (MEC) (Henriksen et al., 2010), which is generally thought to convey almost exclusively spatial information (Hartley et al., 2014). In the framework of our theory, proximal CA1 neurons' functional role may thus be interpreted as combining spatial context (from CA3) with spatial content (via MEC), which may account for the sharper place fields in proximal CA1 (Henriksen et al., 2010). The functional segregation of two spatial inputs is, however, unclear, and we can only speculate that proximal CA1 could be the location to combine context-associated space from CA3 and path-integration-based space from MEC.

A prediction of our model is that distinct principles of information processing are tightly connected with distinct synaptic plasticity mechanisms: The pathway conveying more information in a sparser code should be more plastic than the one conveying less information in a denser code. However, linking this prediction to specific findings in synaptic physiology is difficult. An obvious mechanism enabling distinct forms of plasticity arises from the purely spatial separation into the distinct dendritic compartments with distinct electrotonic properties (Spruston, 2008). For example, the spatially limited backpropagation of dendritic spikes (Golding et al., 2002; Johenning et al., 2009) would allow only proximal synapses to undergo Hebbian-type plasticity. A further way to diversify synaptic plasticity at a single cell is by selective action of neuromodulators. Such effects on LTP and LTD have been reported in both distal and proximal apical synapses connected to several neuromodulatory systems (Otmakhova & Lisman, 1996; Navakkode, Sajikumar, Sacktor, & Frey, 2010; Edelmann & Lessmann, 2011; Nguyen et al., 2014). Most important for this letter, plasticity of LEC inputs can be selectively targeted by dopamine and noradrenaline (Ito & Schuman, 2012).

A hypothesis of particular interest for this letter is that novelty generally enhances synaptic plasticity via the dopaminergic system (Lisman & Otmakhova, 2001; Lisman & Grace, 2005; Ito & Schuman, 2012; McNamara, Tejero-Cantero, Trouche, Campo-Urriza, & Dupret, 2014), which, in our model, would account for synaptic weight changes induced by population patterns $\vec{x}_i$ that are experienced for the first time and require building

a new neural representation in the respective pathway. Importantly, these solitary synaptic changes are not assumed to be minimized in our theory, since they happen only once at the first exposure and require additional energetic effort such as neuromodulatory action.

The sparse hippocampal space code provides a well-investigated test case for our theory; however, it also may be applicable to all other brain regions that are confronted with the need to constantly change associations between (multiple) different input streams. Neocortical layer 5 pyramidal cells, for example, combine bottom-up sensory input stream arriving at proximal dendrites, with top-down feedback at distal dendritic synapses (Larkum, 2013). Our model predicts that the feedback signals that reflect the internal expectations should be encoded densely and relatively nonplastic, whereas the sensory drive should be encoded sparsely (Olshausen & Field, 2004) via plastic synapses.

The mathematics of our model can be generalized from two to multiple (k) input pathways in a straightforward manner. However, in this case, only one pathway (the sparsest one) can remain static, since otherwise a neuron is not able to learn to respond to any of the possible $P_1, \ldots, P_k$ combinations of inputs. Our framework thus potentially applies to many brain areas that require ongoing plastic changes to memorize configurations in a constantly changing world.

Appendix A: Minimizing the Objective Function

Concurrent minimization of weight change and neuron numbers leads to the Lagrange function,

$$\begin{aligned}\mathcal{L}(\kappa_1, \kappa_2, f_1; P_1, P_2, f, K, \lambda) &= \langle \Delta w_1 \rangle^4 + \lambda\,(N_1 + N_2) \\ &= \frac{16\,K\,\sigma^2\,(f_1)^4}{1 - \kappa_1 f_1 - \kappa_2\, f/f_1} + K\,\lambda\,\left(\frac{P_1}{\kappa_1} + \frac{P_2}{\kappa_2}\right),\end{aligned} \tag{A.1}$$

where λ is a Lagrange multiplier that measures the inverse cost of synaptic change. To simplify notation, we introduce

$$c = \sqrt{\frac{16\,\sigma^2 (f_1)^4}{\lambda}}$$

and obtain the minimum of $\mathcal{L}$ for vanishing gradient in $\kappa_{1,2}$:

$$0 = \partial_{\kappa_1} \mathcal{L} \Rightarrow P_1\,(1 - \kappa_1\, f_1 - \kappa_2 f_2)^2 = c^2\, f_1\, \kappa_1^2,$$

$$0 = \partial_{\kappa_2} \mathcal{L} \Rightarrow P_2\,(1 - \kappa_1\, f_1 - \kappa_2 f_2)^2 = c^2\, f_2\, \kappa_2^2.$$

The two resulting equations are solved by

$$\kappa_2 f_2 = \left[1 + \sqrt{\frac{Q_1}{Q_2}} + \frac{c}{\sqrt{Q_2}}\right]^{-1} = \frac{\sqrt{Q_2}}{\sqrt{Q_1} + \sqrt{Q_2} + c} \tag{A.2}$$

and

$$\kappa_1 f_1 = \kappa_2 f_2 \sqrt{\frac{Q_1}{Q_2}} = \frac{\sqrt{Q_1}}{\sqrt{Q_1} + \sqrt{Q_2} + c}. \tag{A.3}$$

Both equations A.2 and A.3 give rise to expressions for the optimal amounts of synapses,

$$N_i = K\left[Q_i + \sqrt{Q_1 Q_2} + c\sqrt{Q_i}\right], \tag{A.4}$$

and, consequently, the minimal cost equals

$$\mathcal{L} = \lambda K \left(\sqrt{Q_1} + \sqrt{Q_2} + c\right)^2 \tag{A.5}$$

A.1 Optimal Sparseness. Inserting equation 2.19 into equation A.5 yields a 1D minimization problem, defined by the reduced-cost function

$$L(f_1)/(K\lambda) = (f_1)^4 \frac{16\sigma^2}{\lambda} + 2\sqrt{\frac{16\sigma^2}{\lambda}} \left(\sqrt{P_1}\,(f_1)^{5/2} + \sqrt{P_2 f}\,(f_1)^{3/2}\right) + P_1 f_1 + P_2 f/f_1 + 2\sqrt{P_1 P_2 f}. \tag{A.6}$$

The polynomial order of $\sqrt{f_1}$ in equation A.6 prohibits a direct analytical solution. Numerical optimization, however, is straightforward.

An approximate solution can be obtained under the precondition that f and f_1 are small. Then only the terms

$$L(f_1)/(K\lambda) \approx P_1 f_1 + 2\sqrt{P_1 P_2 f} + (P_2 f)\,(f_1)^{-1}$$

are of leading order, and the optimal sparseness at which L is minimized equals

$$(f_1)^{\text{opt}} = \sqrt{\frac{f P_2}{P_1}}. \tag{A.7}$$

Because of the sparseness constraints, equation A.7 applies only if $f/\varphi \leq (f_1)^{\text{opt}} \leq \sqrt{f}$; otherwise, $(f_1)^{\text{opt}}$ assumes the upper or lower bound, respectively.

Appendix B: Hippocampus Simulations

For the simulations in Figure 1, we fixed the output sparseness at the approximately observed value of $f = 0.1$ and assume that the sparseness of the object pathway 2 assumes the (information optimal) upper bound $f_2 = 1/2$. The number $P_1 = 750$ of places is a product of the number of environments (30) and the number of places per environment (25). The number P_2 of objects is fixed at 50. According to equation 2.27, these choices yield neuron numbers of $N_1 = 438$ and $N_2 = 179$.

The two input pathways are modeled as follows. The neuronal activity in the place-encoding pathway 1 transmits its information via gaussian place fields; that is, at position $\vec{r}$, the activity of neuron i is modeled as

$$(x_1)_i(\vec{r}) = \zeta_i \exp\left(-\frac{(\vec{r} - \vec{c}_i)^2}{2\sigma^2}\right),$$

with width $\sigma = 1/5$ of the box size and centers $\vec{c}_i$ that uniformly cover the whole 1-by-1 meter box. The random population vector $\vec{\zeta} \in \{0, 1\}^{N_1}$ is chosen such that the number of "1s" is the same for each environment: each environment is encoded by the same number of neurons.

The neuronal activity of neuron j in the object-encoding pathway is modeled as

$$(x_2)_j(\vec{r}, \vec{n}) = \sum_{m=1}^{P_2} \xi_j^m \, n_m \, d(\vec{r}, \vec{r}_m),$$

where $n_m \in \{0, 1\}$ indicates the presence of object m in the present environment at position $\vec{r}_m$ and

$$d(\vec{a}, \vec{b}) = \exp\left(-\frac{(a_x - b_x)^2}{2\sigma_x^2} - \frac{(a_y - b_y)^2}{2\sigma_y^2}\right)$$

introduces a distance-dependent weight with variable-length scales σ_x, σ_y that are uniformly distributed between 1.5 and 2.5 times the box length. The vector $\vec{\xi}^m$ denotes the population pattern in pathway 2 that corresponds to object m. In our simulations, the population patterns were modeled as $\xi_j^m = q_j^m \, \eta_j^m$, where $q \in \{0, 1\}$ is a binary random variable that determines the population sparseness of $p = \text{prob}(q = 1) = 0.25$ and $\eta \in [0, 1]$ is a uniformly distributed random variable. Each environment contains five random

objects at random positions. Since pathway 2 conveys five inputs, Signal_2 is five times as strong as Signal_1, and thus the firing threshold has been set to five times the noise.

The neuronal activity patterns are generated by simulating a random trajectory with constant speed.

Acknowledgments

We are grateful to Álvaro Tejero-Cantero for a critical reading of the manuscript. This work was supported by the German Research Association (DFG) under grant LE-2250/5-1.

References

Cover, T. (1965). Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition. *IEEE Transactions on Electronic Computers*, *14*, 326–334.

Cressant, A., Muller, R. U., & Poucet, B. (1997). Failure of centrally placed objects to control the firing fields of hippocampal place cells. *J. Neurosci.*, *17*, 2531–2542.

Deshmukh, S. S., & Knierim, J. J. (2011). Representation of non-spatial and spatial information in the lateral entorhinal cortex. *Front. Behav. Neurosci.*, *5*, 69.

Edelmann, E., & Lessmann, V. (2011). Dopamine modulates spike timing-dependent plasticity and action potential properties in CA1 pyramidal neurons of acute rat hippocampal slices. *Front. Synaptic Neurosci.*, *3*, 6.

Fusi, S., Drew, P. J., & Abbott, L. F. (2005). Cascade models of synaptically stored memories. *Neuron*, *45*, 599–611.

Fyhn, M., Molden, S., Witter, M. P., Moser, E. I., & Moser, M. B. (2004). Spatial representation in the entorhinal cortex. *Science*, *305*, 1258–1264.

Golding, N. L., Staff, N. P., & Spruston, N. (2002). Dendritic spikes as a mechanism for cooperative long-term potentiation. *Nature*, *418*, 326–331.

Hartley, T., Lever, C., Burgess, N., & O'Keefe, J. (2014). Space in the brain: How the hippocampal formation supports spatial cognition. *Philos. Trans. R. Soc. Lond., B, Biol. Sci.*, *369*, 20120510.

Henriksen, E. J., Colgin, L. L., Barnes, C. A., Witter, M. P., Moser, M. B., & Moser, E. I. (2010). Spatial representation along the proximodistal axis of CA1. *Neuron*, *68*, 127–137.

Hopfield, J. J. (1982). Neural networks and physical systems with emergent collective computational abilities. *Proc. Natl. Acad. Sci. U.S.A.*, *79*, 2554–2558.

Ito, H. T., & Schuman, E. M. (2012). Functional division of hippocampal area CA1 via modulatory gating of entorhinal cortical inputs. *Hippocampus*, *22*, 372–387.

Johenning, F. W., Beed, P. S., Trimbuch, T., Bendels, M. H., Winterer, J., & Schmitz, D. (2009). Dendritic compartment and neuronal output mode determine pathway-specific long-term potentiation in the piriform cortex. *J. Neurosci.*, *29*, 13649–13661.

Kammerer, A., & Leibold, C. (2014). Hippocampal remapping is constrained by sparseness rather than capacity. *PLoS Comput. Biol.*, *10*, e1003986.

Knierim, J. J., Neunuebel, J. P., & Deshmukh, S. S. (2014). Functional correlates of the lateral and medial entorhinal cortex: Objects, path integration and local-global reference frames. *Philos. Trans. R. Soc. Lond., B, Biol. Sci.*, *369*, 20130369.

Larkum, M. (2013). A cellular mechanism for cortical associations: An organizing principle for the cerebral cortex. *Trends Neurosci.*, *36*, 141–151.

Lee, I., Yoganarasimha, D., Rao, G., & Knierim, J. J. (2004). Comparison of population coherence of place cells in hippocampal subfields CA1 and CA3. *Nature*, *430*, 456–459.

Letzkus, J. J., Kampa, B. M., & Stuart, G. J. (2006). Learning rules for spike timing–dependent plasticity depend on dendritic synapse location. *J. Neurosci.*, *26*, 10420–10429.

Leutgeb, S., Leutgeb, J. K., Treves, A., Moser, M. B., & Moser, E. I. (2004). Distinct ensemble codes in hippocampal areas CA3 and CA1. *Science*, *305*, 1295–1298.

Lisman, J. E., & Grace, A. A. (2005). The hippocampal-VTA loop: Controlling the entry of information into long-term memory. *Neuron*, *46*, 703–713.

Lisman, J. E., & Otmakhova, N. A. (2001). Storage, recall, and novelty detection of sequences by the hippocampus: Elaborating on the SOCRATIC model to account for normal and aberrant effects of dopamine. *Hippocampus*, *11*, 551–568.

Lu, L., Leutgeb, J. K., Tsao, A., Henriksen, E. J., Leutgeb, S., Barnes, C. A., . . . Moser, E. I. (2013). Impaired hippocampal rate coding after lesions of the lateral entorhinal cortex. *Nat. Neurosci.*, *16*, 1085–1093.

Mankin, E. A., Sparks, F. T., Slayyeh, B., Sutherland, R. J., Leutgeb, S., & Leutgeb, J. K. (2012). Neuronal code for extended time in the hippocampus. *Proc. Natl. Acad. Sci. U.S.A.*, *109*, 19462–19467.

Mathis, A., Herz, A. V., & Stemmler, M. (2012). Optimal population codes for space: Grid cells outperform place cells. *Neural Comput.*, *24*, 2280–2317.

McNamara, C. G., Tejero-Cantero, A., Trouche, S., Campo-Urriza, N., & Dupret, D. (2014). Dopaminergic neurons promote hippocampal reactivation and spatial memory persistence. *Nat. Neurosci.*, *17*, 1658–1660.

Medan, V., & Preuss, T. (2014). The Mauthner-cell circuit of fish as a model system for startle plasticity. *J. Physiol. Paris*, *108*, 129–140.

Monier, C., Fournier, J., & Fregnac, Y. (2008). In vitro and in vivo measures of evoked excitatory and inhibitory conductance dynamics in sensory cortices. *J. Neurosci. Methods*, *169*, 323–365.

Moser, E. I., Roudi, Y., Witter, M. P., Kentros, C., Bonhoeffer, T., & Moser, M. B. (2014). Grid cells and cortical representation. *Nat. Rev. Neurosci.*, *15*, 466–481.

Navakkode, S., Sajikumar, S., Sacktor, T. C., & Frey, J. U. (2010). Protein kinase Mzeta is essential for the induction and maintenance of dopamine-induced long-term potentiation in apical CA1 dendrites. *Learn. Mem.*, *17*, 605–611.

Neunuebel, J. P, Yoganarasimha, D., Rao, G., & Knierim, J. J. (2013). Conflicts between local and global spatial frameworks dissociate neural representations of the lateral and medial entorhinal cortex. *J. Neurosci.*, *33*, 9246–9258.

Nguyen, C. L., Tran, A. H., Matsumoto, J., Hori, E., Uwano, T., Ono, T., & Nishijo, H. (2014). Hippocampal place cell responses to distal and proximal cue manipulations in dopamine D2 receptor-knockout mice. *Brain Res.*, *1567*, 13–27.

Olshausen, B. A., & Field, D. J. (2004). Sparse coding of sensory inputs. *Curr. Opin. Neurobiol.*, *14*, 481–487.

Otmakhova, N. A., & Lisman, J. E. (1996). D1/D5 dopamine receptor activation increases the magnitude of early long-term potentiation at CA1 hippocampal synapses. *J. Neurosci.*, *16*, 7478–7486.

Päpper, M., Kempter, R., & Leibold, C. (2011). Synaptic tagging, evaluation of memories, and the distal reward problem. *Learn. Mem.*, *18*, 58–70.

Peterson, W., Birdsall, T., & Fox, W. (1954). The theory of signal detectability. *Transactions of the IRE Professional Group on Information Theory*, *4*, 171–212.

Roudi, Y., & Treves, A. (2008). Representing where along with what information in a model of a cortical patch. *PLoS Comput. Biol.*, *4*, e1000012.

Sajikumar, S., & Korte, M. (2011). Different compartments of apical CA1 dendrites have different plasticity thresholds for expressing synaptic tagging and capture. *Learn. Mem.*, *18*, 327–331.

Schaefer, A. T., Larkum, M. E., Sakmann, B., & Roth, A. (2003). Coincidence detection in pyramidal neurons is tuned by their dendritic branching pattern. *J. Neurophysiol.*, *89*, 3143–3154.

Solstad, T., Yousif, H. N., & Sejnowski, T. J. (2014). Place cell rate remapping by CA3 recurrent collaterals. *PLoS Comput. Biol.*, *10*, e1003648.

Spruston, N. (2008). Pyramidal neurons: Dendritic structure and synaptic integration. *Nat. Rev. Neurosci.*, *9*, 206–221.

Torben-Nielsen, B., & Stiefel, K. M. (2010). An inverse approach for elucidating dendritic function. *Front. Comput. Neurosci.*, *4*, 128.

van Rossum, M. C., Shippi, M., & Barrett, A. B. (2012). Soft-bound synaptic plasticity increases storage capacity. *PLoS Comput. Biol.*, *8*, e1002836.

Vazdarjanova, A., & Guzowski, J. F. (2004). Differences in hippocampal neuronal population responses to modifications of an environmental context: Evidence for distinct, yet complementary, functions of CA3 and CA1 ensembles. *J. Neurosci.*, *24*, 6489–6496.

Yoganarasimha, D., Rao, G., & Knierim, J. J. (2011). Lateral entorhinal neurons are not spatially selective in cue-rich environments. *Hippocampus*, *21*, 1363–1374.

Received January 11, 2016; accepted March 20, 2016.

On the representation of space

One of the most important functions of the brain is the development of neural representations of concepts and contexts. Once a particular representation is built, however, it is equally important to be able to convey this information to other brain areas for further processing or a behavioural response. Two main paradigms are used to explain how neuronal activity encodes information: A firing rate code and a spike-time code.

In a firing rate code, average neural activity correlates with the particular features of the world being represented. For instance, the spiking activity of place cells correlates on average with a certain region of space (its place field). But are place cells coding for a particular location within their fields?. Experimental evidence suggests this may be the case. Indeed, it has been proven possible to reconstruct an animal's trajectory from a sequence of activation of a population of place cells, implying they might be specifying individual locations. But how can the brain unambiguously select a desired location from a place field, it being a broad distribution of activity in space?. It turns out a spike-time code might help us finding an answer.

In a spike-time code, information about the represented features is encoded in the precise timing of the spiking activity. Not surprisingly, spike-times are most useful when compared within a population of neurons, where computations at the network level lead to interesting outcomes from the encoded information. A natural reference point are the underlying network rhythms, with the so called Theta rhythm being experimentally linked to place cell spiking activity. With respect to Theta we can frame the spike distributions, which in turn include information about space.

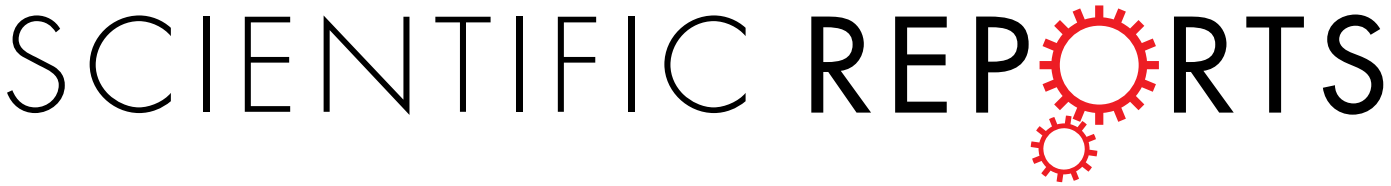
Several questions come now to our minds: How does the shape of the spike distributions affect the selection of the encoded locations?. What are the implications for encoding at the population level?. Are these concerns also shared by the actual biological networks in the Hippocampus?.

In the following paper we develop a framework where the previous questions can be addressed.

All contributions to the following work were done under the supervision of Christian Leibold. C.L. formulated and developed the analytical model. C.L. and M.M. performed the numerical analysis. All authors wrote and reviewed the manuscript.

The results were published in the journal Scientific Reports with the following reference

C. Leibold and M. M. Monsalve-Mercado. 2017. "Traveling Theta Waves and the Hippocampal Phase Code". *Scientific Reports* 7 (1). doi:10.1038/s41598-017-08053-3



OPEN

Traveling Theta Waves and the Hippocampal Phase Code

Christian Leibold[1,2] & Mauro M. Monsalve-Mercado[1,2]

Received: 3 January 2017
Accepted: 6 July 2017
Published online: 09 August 2017

Hippocampal place fields form a neuronal map of the spatial environment. In addition, the distance between two place field centers is proportional to the firing phase difference of two place cells with respect to the local theta rhythm. This consistency between spatial distance and theta phase is generally assumed to result from hippocampal phase precession: The firing phase of a place cell decreases with distance traveled in the place field. The rate of phase precession depends on place field width such that the phase range covered in a traversal of a place field is independent of field width. Width-dependent precession rates, however, generally disrupt the consistency between distance and phase differences. In this paper we provide a mathematical theory suggesting that this consistency can only be secured for different place field widths if phase precession starts at a width-dependent phase offset. These offsets are in accordance with the experimentally observed theta wave traveling from the dorsal to the ventral pole of the hippocampus. Furthermore the theory predicts that sequences of place cells with different widths should be ordered according to the end of the place field. The results also hold for considerably nonlinear phase precession profiles.

Hippocampal place cells fire action potentials (spikes) in only few locations of an environment forming a neuronal map of space[1]. The spike times of place cells are coordinated with the extracellular field potential oscillations in the theta range (4–12 Hz). In experiments with rodents, it was shown that place cells spike late in the theta cycle when the animal enters a place field, and subsequently precess to early theta phases during traversal of the place field[2–5]. As a result of this phase precession, the spike times of a population of place cells arrange as theta sequences that encode trajectories in space[2–4,6–11]: Within each theta cycle cells fire first (at early phases) whose place fields are almost completely traversed, whereas cells fire latest whose place fields have just been entered, and, thus, theta sequences are generally considered as time-compressed representations of the spatial trajectories during behavior. Moreover, theta sequences imply causal pairwise correlations on the theta time scale that can trigger spike-timing-dependent plasticity rules[12–15] and thereby imprint the memories of spatial trajectories into the synaptic matrix of the hippocampal network[4,7,16].

An implicit assumption underlying most of the ideas about decoding theta sequences and their implications for learning is that the rate of phase precession is equal in all neurons and thus accounts for the consistency between contiguous space and the circular theta phase (Fig. 1A); i.e. the phase difference of two neurons is directly proportional to the spatial distance between the positions encoded by the two neurons. However, place cells show different rates of phase precession depending on their place field width[3,8]: Cells with broad place fields precess slower than those with narrow place fields such that the phase range is constant independent of place field width. Owing to the different phase precession rates, spike timing relations between pairs of cells disorganize and cell pairs may swap their order of firing during the course of place field traversal (Fig. 1B).

A further challenge to the assumption that spatial displacements of place fields are consistently encoded as phase differences (called the consistency assumption in what follows) is that the theta oscillation does not constitute a globally synchronized hippocampus-wide oscillation but rather exhibits traveling-wave-like behavior along the dorso-ventral axis[17–20] resulting in a systematic phase shift of the local theta oscillations across the hippocampus. Thus the spikes of more ventral cells, which have larger place fields[21–25] (and precess slowly), generally occur later in time than the spikes of dorsal cells, which have smaller place fields (and precess fast) despite their *local* theta phases might be identical.

Owing to the field-width dependent phase precession rate and the traveling theta wave it is unclear whether the consistency assumption holds across the whole hippocampus, or whether it is confined to only a limited region on the dorso-ventral axis. Since, however, the recurrent connectivity of the hippocampal CA3 network

[1]Department Biology II, Ludwig-Maximilians-Universität München, Munich, Germany. [2]Bernstein Center for Computational Neuroscience Munich, Munich, Germany. Correspondence and requests for materials should be addressed to C.L. (email: leibold@bio.lmu.de)

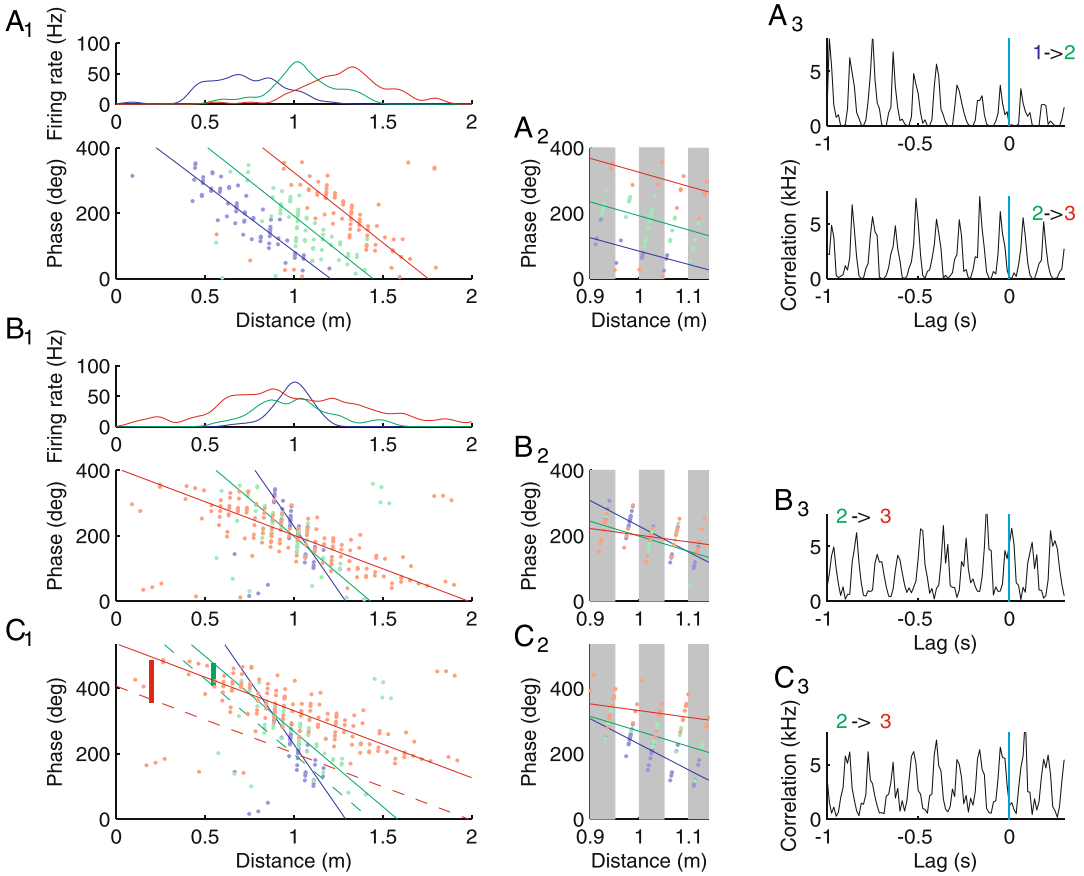


Figure 1. Phase precession and pairwise correlations in simulated place field activity. (**A$_1$**) Spike phases and positions of three cells (colors) with same width and different centers. Top panel depicts firing rates as a function of position. Bottom depicts phase of spikes as a function of position. Solid lines are obtained from circular linear fits to the dots[34]. (**A$_2$**) Close up of phase plot from A$_1$ at the place field center illustrating that in each theta cycle (grey and white columns) the temporal sequence of spikes (vertically increasing phase patterns) corresponds to ordered place fields: 1 before 2 before 3. (**A$_3$**) Crosscorrelation functions exhibit systematic peak shifts on the theta time scale encoding the difference of the place field centers. The negative peak lags correspond to cell 1 firing before cell 2 and cell 2 firing before cell 3 as indicated by the labels ($1 \rightarrow 2$, $2 \rightarrow 3$). (**B$_{1-3}$**) Same as in A for three place cells with different widths and same center. Cells 2 and 3 fire at the same time on average. (**C$_{1-3}$**) Data from B with a cell specific phase shift: The larger the field the more the phases are shifted upwards. Dashed lines are the linear fits from B$_1$. Thick solid vertical lines indicate the theta phase offsets relative to the local theta oscillation of the blue cell: Theta oscillation is increasingly delayed for the green and the red cell. For details on the numerics see Methods.

extends across at least 2/3 of the dorso-ventral axis[26], it is conceivable that the consistency assumption must hold across the whole hippocampus, if the temporal contingency of place field spiking was a fundamental organizational principle of the hippocampal code. So far there is some evidence that parts of the trajectories can indeed be decoded from theta sequences across large parts of the dorso-ventral extent of the hippocampus[20], however, a consistent theoretical framework is missing.

In this paper we provide a theoretical analysis of the consistency assumption taking into account variable place-field-width-dependent rates of phase precession. Our analysis shows that the consistency assumption can indeed be secured in a traveling wave framework if cells with broader place fields phase precess with respect to a local theta rhythm that is delayed compared to the local theta rhythm for cells with smaller place fields (Fig. 1C). From an optimality argument, we will derive that the maximal phase shift between the hippocampal theta oscillations in the dorsal and the ventral hippocampus should be about 180° as was found experimentally[18,19].

Methods

Numerical Simulations. For Fig. 1 we simulated place field activity as inhomogeneous Poisson processes with density

$$\lambda(t) \propto \exp\left[-\frac{(vt - x_0)^2}{2\sigma^2}\right][1 + \cos(\omega_c(t - x_0/v))]^4 \quad (1)$$

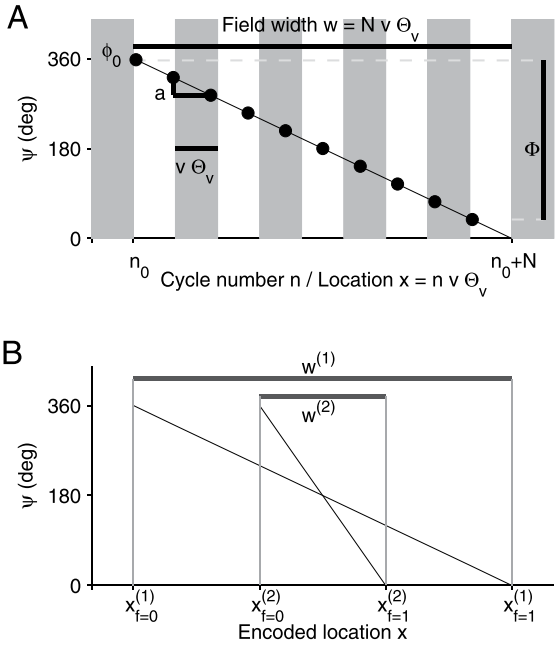


Figure 2. Nomenclature. (**A**) Spike phases ψ (black circles) decrease by a in each theta cycle (alternating grey and white patches). A field starts at theta cycle n_0 and ends at cycle $n_0 + N$ (here $N = 9$). The start phase is denoted ϕ_0, the total phase range is denoted by $\Phi < 0$. The width w of the field is a result of the number N of theta cycles needed to cover the phase range. The spatial distance an animal covers in one theta cycle equals $v\Theta_v$, in which v denotes running speed and Θ_v denotes the period length of a theta cycle in time. (**B**) To define a ordered sequence of overlapping place fields of different widths, we introduce the parameter f (fraction in the field), which identifies the field's location in the sequence. For $f = 0$ (beginning of the field) the proposed sequential ordering of the two fields is $1 \rightarrow 2$, for $f = 1$ (end of the field) the ordering is $2 \rightarrow 1$.

were v denotes running speed, t is time, x_0 the place field center, σ specifies place field width, and $\omega_c = 2\pi/T_c$ is the oscillation frequency of the individual neurons. In all simulations we assumed a linear path with constant speed $v = 40$ cm/s. The oscillation period of a cell was width dependent to ensure a width independent phase range, $T_c = T_\theta\left(1 - 0.06\frac{\sigma_0}{\sigma}\right)$, with $\sigma_0 = 20$ cm and the theta period $T_\theta = 1/8$ s.

Results

Model. We assume that within a place field the theta phase ψ of the spikes decreases linearly by a constant amount a per cycle (see Fig. 2A for illustration). If the animal enters the place field of cell i at theta cycle $n_0^{(i)}$, the phase $\psi_n^{(i)}$ of neuron i at cycle n is thus described by

$$\psi_n^{(i)} = \phi_0^{(i)} - a(v, w^{(i)})(n - n_0^{(i)}),\ n \geq n_0^{(i)}. \tag{2}$$

The phase offset $\phi_0^{(i)}$ denotes the starting phase at the entrance of the field. Most importantly, this offset is cell specific (indicated by i) and thus the local theta rhythm is not introducing a hippocampus-wide temporal reference. Later on, we will relate $\phi_0^{(i)}$ to the hippocampus-wide temporal reference frame introduced by the travelling theta wave. The slope parameter $a > 0$ in equation (2) depends on both the running speed v of the animal and the width $w^{(i)}$ of the place field, defined as the spatial distance between the occurrence of the first and the last theta cycle. The slope is such that the phase range Φ that is covered while fully crossing a place field during $N^{(i)}$ theta cycles is independent of speed[27,28] and width[8,27,29], and identical for all cells, i.e.,

$$\psi^{(i)}_{n_0^{(i)}+N^{(i)}} - \psi^{(i)}_{n_0^{(i)}} = \Phi \Rightarrow -a(v, w^{(i)})N^{(i)} = \Phi. \tag{3}$$

Experimental reports restrict the phase range to values below 2π[5,30,31], and therefore we can treat the phases ψ as linear variables without the risk of potential ambiguities.

The place field width $w^{(i)}$ can be expressed in terms of the speed-dependent theta period Θ_v and the running speed v according to

$$w^{(i)} = v\Theta_v N^{(i)} \tag{4}$$

and thus the slope parameter equals

$$a(v, w^{(i)}) = \frac{-\Phi v\Theta_v}{w^{(i)}}. \tag{5}$$

Coding Assumptions. Place field activity encodes a spatial position, however, it is not a priori clear what exact position this should be since place fields are extended in space. For the rate code, the place field is generally interpreted in a probabilistic way in that the firing rate is seen as a correlate for the probability of being at a certain position. This argument, however, does not pertain to a timing code. We thus have to make additional assumptions about the encoding of space by the theta phase. In particular, thinking about sequences of place cells we have to deal with the problem of which of the two place fields comes earlier in the sequence. Since place fields have different widths, this sequence will be generally different depending on whether we order the fields according to their beginning, their center, or their end. In the following we assume that the timing of a place cell spike encodes the distance to the sequence position

$$x^{(i)} := (n_0^{(i)} + n_x^{(i)})v\,\Theta_v,\ n_x^{(i)} = fN^{(i)}, \tag{6}$$

i.e., the distance to the position that corresponds to the fraction f of the run through the place field. Choosing $f=0$ would mean that a spike encodes the distance from the beginning of the place field, $f=1/2$ would mean that a spike encodes the distance to middle of the field and $f=1$ indicates the distance to the end of the field (see Fig. 2B for illustration).

Phase Difference. Equations (4) and (6) let us relate the starting cycle $n_0^{(i)}$ to the reference position $x^{(i)}$ via

$$n_0^{(i)} = \frac{x^{(i)} - fw^{(i)}}{v\Theta_v}, \tag{7}$$

and thus, combining eqs (2), (5) and (7), we can express the phase difference between two cells i and j as

$$\begin{aligned}\Delta^{(ij)}\psi_n \;&:=\; \psi_n^{(i)} - \psi_n^{(j)} = \Delta^{(ij)}\phi_0 + \Phi v\Theta_v\left[\frac{n-n_0^{(i)}}{w^{(i)}} - \frac{n-n_0^{(j)}}{w^{(j)}}\right]\\ &= \begin{cases}\Delta^{(ij)}\phi_0 - \dfrac{\Phi\Delta^{(ij)}x}{w^{(i)}} + \dfrac{\Phi\Delta^{(ij)}w}{w^{(i)}}\left[f - \dfrac{v\Theta_v}{w^{(j)}}(n-n_0^{(j)})\right] & \text{if } n_0^{(j)} \geq n_0^{(i)}\\ \Delta^{(ij)}\phi_0 - \dfrac{\Phi\Delta^{(ij)}x}{w^{(j)}} + \dfrac{\Phi\Delta^{(ij)}w}{w^{(j)}}\left[f - \dfrac{v\Theta_v}{w^{(i)}}(n-n_0^{(i)})\right] & \text{if } n_0^{(i)} \geq n_0^{(j)}\end{cases},\end{aligned} \tag{8}$$

where in general we denote differences by

$$a^{(i)} - a^{(j)} := \Delta^{(ij)}a$$

and expanded $n - n_0^{(i)} = n - n_0^{(j)} + n_0^{(j)} - n_0^{(i)}$ for $n_0^{(j)} \geq n_0^{(i)}$, and $n - n_0^{(j)} = n - n_0^{(i)} + n_0^{(i)} - n_0^{(j)}$ for $n_0^{(i)} \geq n_0^{(j)}$.

Special case $w^{(i)} = w^{(j)} = w$. Most coding ideas regarding phase precession implicitly assume place fields of equal width. In such a case equation (8) simplifies to

$$\Delta^{(ij)}\psi_n = \Delta^{(ij)}\phi_0 - \frac{\Phi\Delta^{(ij)}x}{w}. \tag{9}$$

A consistent phase code requires that $\Delta^{(ij)}\psi_n = 0$ for $\Delta^{(ij)}x = 0$, and thus the difference in phase offsets must vanish,

$$\Delta^{(ij)}\phi_0 = 0, \tag{10}$$

i.e., all neurons should start the place field firing at the same phase ϕ_0. Therefore, when the two cells encode different positions, we have

$$\Delta^{(ij)}\psi = -\frac{\Phi}{w}\Delta^{(ij)}x. \tag{11}$$

The phase shift is therefore proportional to the place field difference consistent with the experimentally reported phase code for distance[8,9,32] and illustrated in Fig. 1A.

General case $w^{(i)} \neq w^{(j)}$. The most obvious difference to the special case of equal widths is that, for unequal place field widths $w^{(i)} \neq w^{(j)}$, where the phase difference depends on the cycle number n, the phase relation changes with time; see equation (8) and Fig. 1B. Consistency between the phase and the place code thus cannot be achieved on a cycle by cycle basis, however, the consistency argument can be generalized if one assumes that for cells with $\Delta^{(ij)}x = 0$, the phase difference has to be zero averaged over $N+1$ spike pairs in N theta cycles. Averaging is a biologically plausible computation, since spike-timing dependent synaptic learning rules[15] that encode sequence memories are able to average over multiple repetitions of spike pairs, which in the present case would be averaging over all theta cycles in a place field traversal[4,7,16].

To compute the average phase difference $\langle\Delta^{(ij)}\psi\rangle = (N^{(j)}+1)^{-1}\sum_n \Delta^{(ij)}\psi_n$, we assume without loss of generality that $w^{(i)} > w^{(j)}$, and $n_0^{(j)} \geq n_0^{(i)}$, and consequently $n_0^{(j)} + N^{(j)} \leq n_0^{(i)} + N^{(i)}$, because if the narrower field (j) would

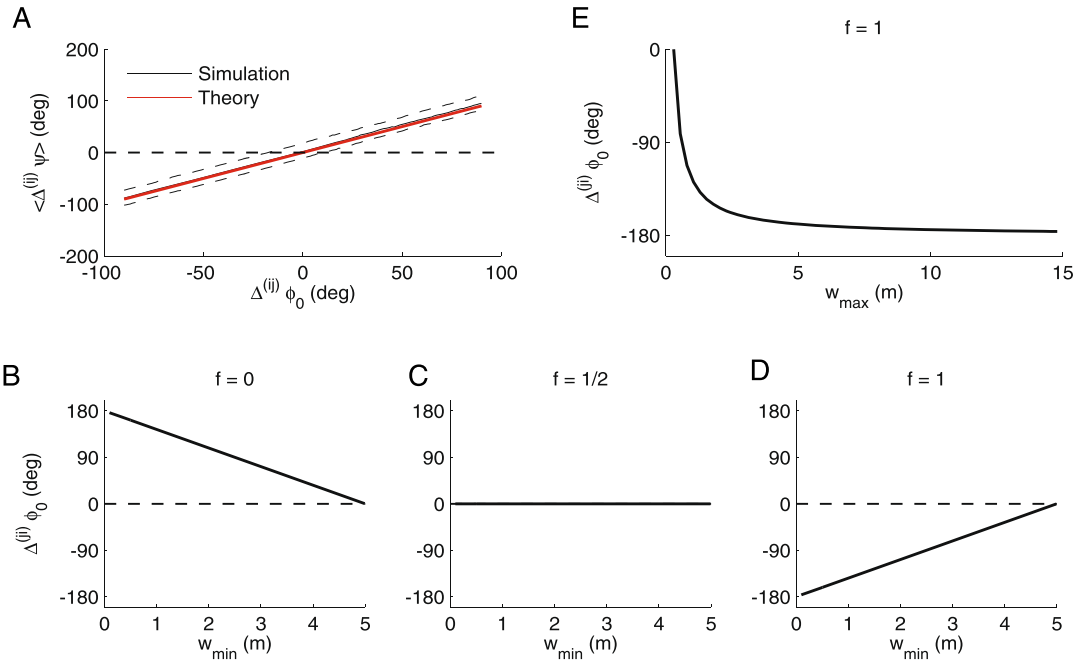


Figure 3. Optimal phase offsets. (**A**) To validate the theoretical results (red line) from equation (12), we derived empirical mean phase differences from simulations (see Fig. 1) of a pair of phase precessing place cells with the same center. Black line depicts the median of 50 repetitions (dashed lines are the 10- and 90-percentiles). The only free parameter was the phase range which we estimated as $\Phi = -360°$ from the simulations. In the example we used $f = 1/2$ and $v = 0.2$ m/s. (**B**–**D**) Optimal phase offset from equation (13) as a function of the width of the smaller place field, when the larger place field width was taken to be 5 m. (**E**) Optimal phase offset as a function of the larger field width while fixing the smaller field width to 30 cm.

start before or end after the wider (i) field, the two cells would necessarily not encode for the same position. Under these conditions we obtain

$$\begin{aligned}\sum_{n=n_0^{(j)}}^{n_0^{(j)}+N^{(j)}} \Delta^{(ij)}\psi_n &= (N^{(j)}+1)\left[\Delta^{(ij)}\phi_0 + \frac{\Delta^{(ij)}w}{w^{(i)}}\Phi f\right] - \frac{\Delta^{(ij)}w}{w^{(i)}w^{(j)}}\Phi v\Theta_v\frac{N^{(j)}(N^{(j)}+1)}{2} \\ &= (N^{(j)}+1)\left[\Delta^{(ij)}\phi_0 + \frac{\Delta^{(ij)}w}{w^{(i)}}\Phi\left(f-\frac{1}{2}\right)\right].\end{aligned} \tag{12}$$

Consequently, the offset difference $\Delta^{(ij)}\phi_0$ for which the average phase difference from equation (12) vanishes is given by

$$\Delta^{(ij)}\phi_0 = -\frac{\Phi\Delta^{(ij)}w}{2w^{(i)}}(2f-1). \tag{13}$$

Comparison of the average phase difference at the theoretically optimal phase offset $\Delta^{(ij)}\phi_0$ from equation (13) agrees with those from a simulated pair of phase precessing cells (Fig. 3A). A biological interpretation of the fundamental equation (13) will be found in the next section.

Biological Constraints. *Phase Range.* At first, we were asking how the theoretical phase offset $\Delta^{(ij)}\phi_0$ from equation (13) fits to the observed theta wave traveling from the dorsal to the ventral pole[17–19]. In the framework of our theory a traveling wave would account for different phase offset ϕ_0 at different dorso-ventral positions. The difference in phase offset $\Delta^{(ij)}\phi_0$ can thus potentially be interpreted as a wave traveling from place cells of width $w^{(i)}$ to place cells with width $w^{(j)}$. We therefore computed $\Delta^{(ji)}\phi_0 = -\Delta^{(ij)}\phi_0$ for a pair of place fields as a function of the width $w_{\min}$ of the smaller place field for changing fractions f that determine the reference position that the spike phase is supposed to encode the distance from (Fig. 3B–D). For $f = 0$ (spike timing encodes distance from the beginning of the field) the optimal phase offset $\Delta^{(ji)}\phi_0$ for small fields is positive (delayed) corresponding to a wave traveling from large place fields to small place fields. For $f = 1/2$ (spike timing encodes distance from the center of the field) the optimal phase offset is mostly close to zero (synchronous) corresponding to a global oscillation. Finally, for $f = 1$ (spike timing encodes the distance to the end of the field) the optimal phase offset for small fields is negative (advanced) corresponding to a wave traveling from small place fields to large place fields as it would be consistent with the observed traveling direction of the theta wave. The maximum phase offset (for the smallest place fields) is the 180° (Fig. 3E) found in experiments[18,19].

Our work thus suggests that the theta phase offset between dorsal and ventral pole is in fact bringing the place cell spikes together in time such that neighboring cells can be encoded by downstream coincidence detector neurons irrespective of place cell width. The fundamental open question is "What are neighbors?". This question is essentially reflected by the parameter f, which says that cells are neighbors if their reference positions $x = x_0 + fw$ are neighboring. Our results show that the traveling wave from dorsal to ventral pole can bring neighboring cells only into temporal coincidence for $f = 1$.

Traveling Waves. Mathematically the phase $\varphi(x_w, t)$ of a wave can be written as

$$\varphi(x_w, t, v) = k_{x_w,v} x_w - t\omega_v + \alpha_v \tag{14}$$

where $k_{x_w,v}$ is the wave vector (2π divided by wavelength), x_w is the position along the dorso-ventral axis, $\omega_v = (2\pi)/\Theta_v$ is the speed-dependent theta oscillation frequency, and α_v is some (arbitrary) constant phase. Equaling the wave phase equation (14) with the optimal phase offset $\Delta^{(ij)}\phi_0$ at the ventral most pole ($w_{\max} = \max_i w_i$) from equation (13) with $w = w^{(j)}$, and $f = 1$ yields

$$\alpha_v + k_{x_w,v} x_w = \phi_0^{(i)} - \phi_0^{(j)} = -\frac{\Phi}{2} + \frac{\Phi w}{2w_{\max}}, \tag{15}$$

where without loss of generality we set the reference phase at the ventral pole, i.e., $\varphi(w_{\max}, t, v) = -t\omega_v$. Since field width increases along the dorso-ventral axis, we can assume $x_w \propto w/w_{\max}$ and thus obtain the optimal wave vector to be constant,

$$k_{x_w,v} = k. \tag{16}$$

The remaining term in eq. (15) can be identified with a constant phase $\alpha_v = -\Phi/2$. The wave propagation speed $c(v)$ is the time derivative of a position $x_\varphi(t)$ of constant phase φ. Since

$$\varphi = kx_\varphi(t) - \omega_v t + \alpha, \tag{17}$$

taking the time derivative yields

$$c(v) \equiv \frac{\mathrm{d}}{\mathrm{d}t} x_\phi = \frac{\omega_v}{k}, \tag{18}$$

which is a function of speed (v), because theta frequency (ω_v) changes with speed. To obtain an explicit expression we assume that space x and place field width w are related by

$$x_w = X\frac{w}{w_{\max}}, \tag{19}$$

where X quantifies the total spatial extent of the dorso-ventral axis of about 1 cm in rats[19]. Combining equations (15), (18) and (19), we obtain

$$c(v) = \frac{4\pi X/|\Phi|}{\Theta_v}. \tag{20}$$

For a phase range of $|\Phi| = 2\pi$, equation (20) yields a velocity estimate of

$$c(v) = \frac{2X}{\Theta_v}. \tag{21}$$

For approximate values of $X = 1$ cm and $1/\Theta_v \approx 8$ Hz this amounts to a propagation velocity of $c \approx 16$ cm/s consistent with experiments[19].

Nonlinear phase precession. The specific shape of phase precession may deviate from the linear model assumption[4,30]. We therefore asked how much non-linear precession would affect the conclusions from the linear model. To parameterize the non-linearity we replace the linear term $\phi_0 + \Phi\frac{n - n_0^{(i)}}{N^{(i)}}$ from eq. (2) by $(\phi_0 + \Phi) - \Phi\left(1 - \frac{n - n_0^{(i)}}{N^{(i)}}\right)^\mu$ with some positive exponent $0 < \mu < 1$. The smaller μ the more non-linear the phase dependence becomes. Following a similar derivation as described for the linear case we end up at an optimal phase offset of

$$\Delta^{ij}\phi_0 = \frac{\Phi}{N^{(j)} + 1}\sum_{m=0}^{N^{(j)}}\left[\left(1 - \frac{m}{N^{(i)}} - f\frac{\Delta^{(ij)}w}{w^{(i)}}\right)^\mu - \left(1 - \frac{m}{N^{(j)}}\right)^\mu\right]. \tag{22}$$

The optimal phase offset from eq. (22) can be numerically evaluated as shown in Fig. 4.

If the exponent μ is below the value of 1, we observe two main effects. First, the optimal phase offset becomes dependent on running speed v, however, only for large differences in place field width. Second, the optimal phase offsets stay below 180°. If the largest field width is fixed (the reference for theta phase), both speed dependence

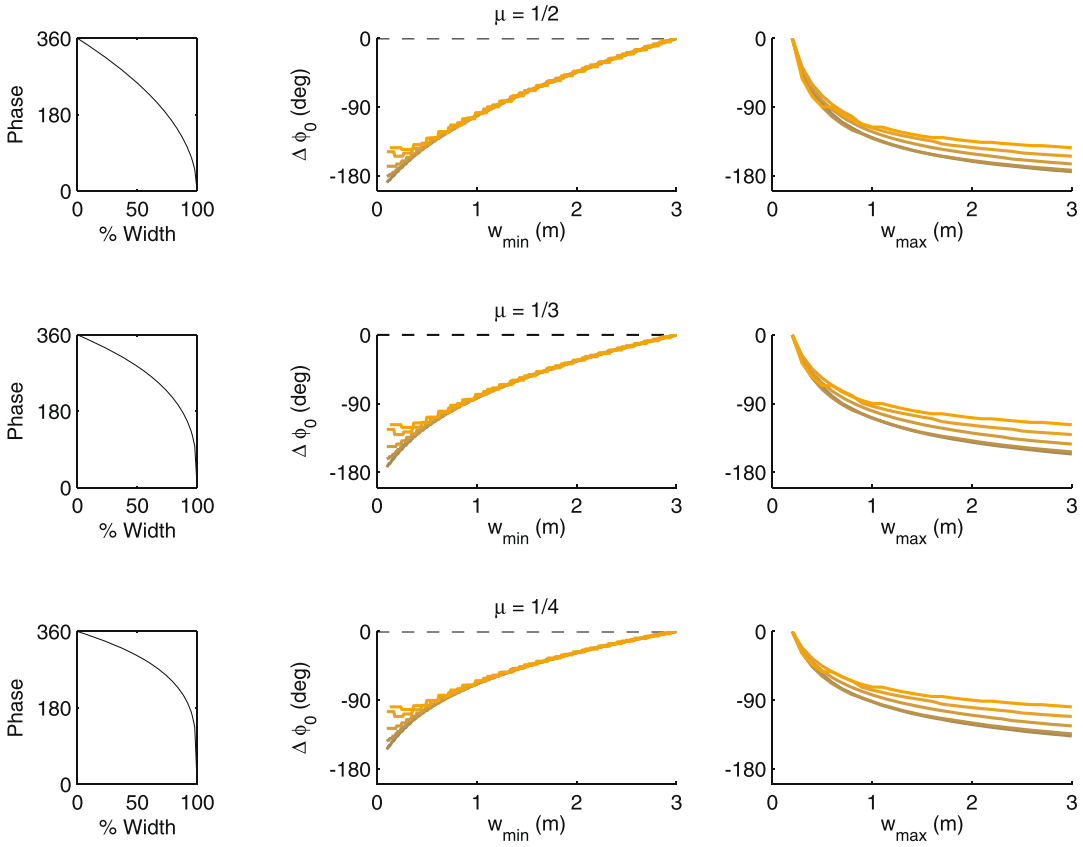


Figure 4. Nonlinear phase precession. (Left column) Non-linear models of phase precession with increasing curvature from top to bottom. (Middle) Optimal phase range for fixed maximal width and $f = 1$ (as in Fig. 2D). The different colors indicate different running speeds from 0.1 m/s (dark) to 1.2 m/s (bright). (Right) Optimal phase range for fixed minimal field width (as in Fig. 2E).

and phases below 180° can be observed for small place field widths below about 0.25 m; smaller than 180° phase offsets are found for high velocities. If, in contrast, the small field width (dorsal pole) is taken as a reference, the velocity dependence extends over a larger range of place field sizes. Also the curves are generally not linear indicating a width-dependent conduction speed of the optimal wave that is faster towards the dorsal end (corresponding to the steeper slopes).

If the non-linearity is moderate ($\mu \gtrapprox 1/2$) the optimal phase offsets can nevertheless still be reconciled with the observed traveling wave, particularly considering that the consistency assumption is mostly violated for high running speeds and low place field widths, where only few theta cycles contribute to potential synaptic weight changes. However, for strong non-linearities ($\mu \lessapprox 1/4$), the reduced phase range questions the validity of the consistency assumption. We thus conclude that the consistency assumption is quite robust for moderately non-linear phase precession and larger place field width.

Discussion

The precise timing of hippocampal place cell firing relative to the local theta oscillation contains information about the position of an animal. Particularly it has been proposed that spatial distances are encoded by hippocampal theta phase differences. In this paper, we theoretically evaluated under which conditions this hypothesis holds. We found that despite the variable place field sizes (and hence variable precession rates), consistency between phase code and spatial distance of place fields is (approximately) possible because the hippocampal theta oscillation is associated with a traveling wave that moves from the dorsal to the ventral pole. The traveling wave thereby imposes a location-specific phase offset, which delays the spikes of the more ventral neurons in time and, in so doing, accounts for the consistency between space and phase differences on average. Our model predicts that the maximal theta phase offset between dorsal and ventral pole should be about 180 degrees as found in experiments[18,19].

A direct consequence of our theory is that it predicts that the reference positions to which distances are encoded by the theta phase of spikes must be the ends of the place fields, otherwise the consistency assumption could not be reconciled with the direction of the traveling wave. As a result action potentials would encode positions the animal would reach in the future and thus, this prediction is consistent with the previously proposed hypothesis that theta sequences predict future behaviors[6,8,10]. Rate-based theories of the hippocampal place code generally assume that place cell activity is linked to the current position of the animal. While this is a perfect assumption to optimally reconstruct animal trajectories from neural activity, our results, however, argue for the

development of predictive strategies for decoding hippocampal place cell activity that optimize estimates of the future trajectory of the animal.

Although place field widths generally increase along the dorso-ventral axis, there is considerable variability at each location[21–25], which may pose a problem to the presented theory. However, some hippocampal regions show more variability than others. For example the correlation between size and location seems the strongest in area CA3[22], which would make this region the most likely candidate area to look at. Also the field potential oscillation is an average over the synaptic inputs of many cells and thus it cannot be excluded that individual cells at similar dorso-ventral positions show distinct intracellular theta phases. Our theory therefore predicts that for place fields with different widths but at similar dorso-ventral position (seeing the same local theta phase), the offset of the theta phase precession should be correlated with field width. Broader place fields (with shallower precession slopes) should start firing at later theta phases to ensure consistency between phase and place field distance.

Theta phase precession and spike correlations (theta sequences) have been argued to be to some degree distinct phenomena of spike time coordination[33], i.e., theta sequences can be altered without observable changes in phase precession. This dissociation shows that the coordination of phase offsets across cells plays an important role in shaping the hippocampal ensemble code, and that the two key features, phase slope and offset, may rely on two distinct mechanisms. Particularly CA3 seems to be fundamental in coordinating the offsets, since inactivation of CA3 removes theta sequences while leaving phase precession intact on the single cell level[33]. Conversely, the phase precession slopes seem to depend at least partly on the medial entorhinal cortex (MEC) since animals with lesioned MEC displayed very little single cell phase precession[32].

References

1. O'Keefe, J. Place units in the hippocampus of the freely moving rat. *Exp Neurol* **51**, 78–109 (1976).
2. O'Keefe, J. & Recce, M. L. Phase relationship between hippocampal place units and the EEG theta rhythm. *Hippocampus* **3**, 317–330, doi:10.1002/hipo.450030307 (1993).
3. Skaggs, W. E., McNaughton, B. L., Wilson, M. A. & Barnes, C. A. Theta phase precession in hippocampal neuronal populations and the compression of temporal sequences. *Hippocampus* **6**, 149–172, doi:gt;3.0.CO;2-K (1996).
4. Mehta, M. R., Lee, A. K. & Wilson, M. A. Role of experience and oscillations in transforming a rate code into a temporal code. *Nature* **417**, 741–746, doi:10.1038/nature00807 (2002).
5. Schmidt, R. *et al*. Single-trial phase precession in the hippocampus. *J Neurosci* **29**, 13232–13241, doi:10.1523/JNEUROSCI.2270-09.2009 (2009).
6. Jensen, O. & Lisman, J. E. Hippocampal CA3 region predicts memory sequences: accounting for the phase precession of place cells. *Learn. Mem.* **3**, 279–287 (1996).
7. Melamed, O., Gerstner, W., Maass, W., Tsodyks, M. & Markram, H. Coding and learning of behavioral sequences. *Trends Neurosci.* **27**, 11–14 (2004).
8. Dragoi, G. & Buzsáki, G. Temporal encoding of place sequences by hippocampal cell assemblies. *Neuron* **50**, 145–157, doi:10.1016/j.neuron.2006.02.023 (2006).
9. Geisler, C. *et al*. Temporal delays among place cells determine the frequency of population theta oscillations in the hippocampus. *Proc Natl Acad Sci USA* **107**, 7957–7962, doi:10.1073/pnas.0912478107 (2010).
10. Feng, T., Silva, D. & Foster, D. J. Dissociation between the experience-dependent development of hippocampal theta sequences and single-trial phase precession. *J. Neurosci.* **35**, 4890–4902 (2015).
11. Wang, Y., Romani, S., Lustig, B., Leonardo, A. & Pastalkova, E. Theta sequences are essential for internally generated hippocampal firing fields. *Nat. Neurosci.* **18**, 282–288 (2015).
12. Gerstner, W., Kempter, R., van Hemmen, J. L. & Wagner, H. A neuronal learning rule for sub-millisecond temporal coding. *Nature* **383**, 76–81 (1996).
13. Markram, H., Lubke, J., Frotscher, M. & Sakmann, B. Regulation of synaptic efficacy by coincidence of postsynaptic APs and EPSPs. *Science* **275**, 213–215 (1997).
14. Kempter, R., Gerstner, W. & van Hemmen, J. L. Hebbian learning and spiking neurons. *Phys. Rev. E: Stat., Nonlin., Soft Matter Phys.* **59**, 4498–4514 (1999).
15. Bi, G. & Poo, M. Distributed synaptic modification in neural networks induced by patterned stimulation. *Nature* **401**, 792–796 (1999).
16. Byrnes, S., Burkitt, A. N., Grayden, D. B. & Meffin, H. Learning a sparse code for temporal sequences using STDP and sequence compression. *Neural Comput* **23**, 2567–2598 (2011).
17. Lubenov, E. V. & Siapas, A. G. Hippocampal theta oscillations are travelling waves. *Nature* **459**, 534–539 (2009).
18. Royer, S., Sirota, A., Patel, J. & Buzsaki, G. Distinct representations and theta dynamics in dorsal and ventral hippocampus. *J. Neurosci.* **30**, 1777–1787 (2010).
19. Patel, J., Fujisawa, S., Berenyi, A., Royer, S. & Buzsaki, G. Traveling theta waves along the entire septotemporal axis of the hippocampus. *Neuron* **75**, 410–417 (2012).
20. McKenzie, S. A., Patel, J., Chadwick, A. & Buzsáki, G. Segmentation of space by theta sequences recorded across the longitudinal axis of the hippocampus. *SfN Poster Abstract* **257**, 13/X47 (2015).
21. Jung, M. W., Wiener, S. I. & McNaughton, B. L. Comparison of spatial firing characteristics of units in dorsal and ventral hippocampus of the rat. *J. Neurosci.* **14**, 7347–7356 (1994).
22. Kjelstrup, K. B. *et al*. Finite scale of spatial representation in the hippocampus. *Science* **321**, 140–143 (2008).
23. Maurer, A. P., Vanrhoads, S. R., Sutherland, G. R., Lipa, P. & McNaughton, B. L. Self-motion and the origin of differential spatial scaling along the septo-temporal axis of the hippocampus. *Hippocampus* **15**, 841–852 (2005).
24. Ahmed, O. J. & Mehta, M. R. The hippocampal rate code: anatomy, physiology and theory. *Trends Neurosci.* **32**, 329–338 (2009).
25. Henriksen, E. J. *et al*. Spatial representation along the proximodistal axis of CA1. *Neuron* **68**, 127–137 (2010).
26. Wittner, L., Henze, D. A., Zaborszky, L. & Buzsaki, G. Three-dimensional reconstruction of the axon arbor of a CA3 pyramidal cell recorded and filled *in vivo*. *Brain Struct Funct* **212**, 75–83 (2007).
27. Huxter, J., Burgess, N. & O'Keefe, J. Independent rate and temporal coding in hippocampal pyramidal cells. *Nature* **425**, 828–832, doi:10.1038/nature02058 (2003).
28. Geisler, C., Robbe, D., Zugaro, M., Sirota, A. & Buzsáki, G. Hippocampal place cell assemblies are speed-controlled oscillators. *Proc Natl Acad Sci USA* **104**, 8149–8154, doi:10.1073/pnas.0610121104 (2007).
29. Terrazas, A. *et al*. Self-motion and the hippocampal spatial metric. *J. Neurosci.* **25**, 8085–8096 (2005).
30. Yamaguchi, Y., Aota, Y., McNaughton, B. L. & Lipa, P. Bimodality of theta phase precession in hippocampal place cells in freely running rats. *J. Neurophysiol.* **87**, 2629–2642 (2002).
31. Harris, K. D. *et al*. Spike train dynamics predicts theta-related phase precession in hippocampal pyramidal cells. *Nature* **417**, 738–741 (2002).

32. Schlesiger, M. I. *et al*. The medial entorhinal cortex is necessary for temporal organization of hippocampal neuronal activity. *Nat. Neurosci.* **18**, 1123–1132 (2015).
33. Middleton, S. J. & McHugh, T. J. Silencing CA3 disrupts temporal coding in the CA1 ensemble. *Nat. Neurosci.* **19**, 945–951 (2016).
34. Kempter, R., Leibold, C., Buzsaki, G., Diba, K. & Schmidt, R. Quantifying circular-linear associations: hippocampal phase precession. *J. Neurosci. Methods* **207**, 113–124 (2012).

Acknowledgements

We would like to thank Jill Leutgeb and Anton Sirota for discussions. This work was funded by the Deutsche Forschungsgemeinschaft (DFG) under grant number LE2250/5-1.

Author Contributions

C.L. formulated the model. C.L. and M.M. performed the numerical analysis. All authors wrote and reviewed the manuscript.

Additional Information

Competing Interests: The authors declare that they have no competing interests.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

On grid cell firing patterns

The discovery of grid cells more than two decades after knowing about the existence of place cells marked a turning point for the understanding of how space is represented in the brain. Since then, much focus have been devoted to explaining the reasons behind the appearance of the characteristic hexagonal firing patterns.

Hexagonal patterns are found abundantly in nature. Their repeated occurrence in nature is due to the fundamental fact that the hexagonal lattice is the most optimal arrangement to densely fill two-dimensional space. Examples of natural systems making use of this fact range from many simple structures such as stable molecules and materials like graphene to more complex physical and biological systems. The organisation of soap bubbles, the cells of a bee honeycomb, the patterns in Rayleigh-Bénard convection, the wavefunction of Bose-Einstein condensates, and some chemical reaction-diffusion systems are all examples of the emergence of hexagonal patterns in nature. But could a complex biological system, as illustrated by the spiking activity of grid cells, be the result of following such simple principles?.

In the following letter we explore the biological mechanisms by which the hexagonal firing patterns of grid cells emerge. We highlight the interaction between the entorhinal cortex and hippocampus proper as the driving force behind the formation of the peculiar hexagonal patterns. In particular, the place cell input coming from areas CA1 and subiculum provides the entorhinal cortex with information about space in a special way, such that its processing leads to the implementation of the simple principles of dense filling.

The key biological mechanism underlying this process is a generalisation of Hebbian plasticity, which takes into account the precise timing of the input spiking activity to transform the encoded information about space into the distinct grid cell firing patterns. Next, we introduce how these plastic changes are motivated and develop an analytical framework where the relation to self-organising patterns becomes evident.

All contributions to the following work were done under the supervision of Christian Leibold. M.M. formulated and developed the analytical model and performed the numerical analysis. All authors wrote and reviewed the manuscript.

Preliminary results of this work were presented as a poster in the Bernstein Conference 2016 in Berlin with the following reference

Monsalve Mercado MM, Leibold C (2016). Hippocampal phase precession leads to grid cells via STDP learning. Bernstein Conference 2016. doi: 10.12751/nncn.bc2016.0202

The final results were presented as a selected talk at the 3rd International Conference on Mathematical Neuroscience in Boulder, CO. The talk contribution can be found as

Monsalve Mercado MM, Leibold C (2017). Feed-forward learning of grid cells: Universality and biological implementations. ICMNS 2017, Boulder, CO, USA.

The results were published in the journal Physical Review Letters with the following reference

M. M. Monsalve-Mercado and C. Leibold. 2017. "Hippocampal Spike-Timing Correlations Lead to Hexagonal Grid Fields". *Physical Review Letters* 119, no. 3 (): 38101. ISSN: 0031-9007. doi:10.1103/PhysRevLett.119.038101

Hippocampal Spike-Timing Correlations Lead to Hexagonal Grid Fields

Mauro M. Monsalve-Mercado and Christian Leibold
Department Biologie II & Graduate School of Systemic Neurosciences, LMU Munich, Großhadernerstr. 2, 82152 Planegg, Germany, and Bernstein Center for Computational Neuroscience Munich, Großhadernerstr. 2, 82152 Planegg, Germany

(Received 10 February 2017; revised manuscript received 16 May 2017; published 19 July 2017)

Space is represented in the mammalian brain by the activity of hippocampal place cells, as well as in their spike-timing correlations. Here, we propose a theory for how this temporal code is transformed to spatial firing rate patterns via spike-timing-dependent synaptic plasticity. The resulting dynamics of synaptic weights resembles well-known pattern formation models in which a lateral inhibition mechanism gives rise to a Turing instability. We identify parameter regimes in which hexagonal firing patterns develop as they have been found in medial entorhinal cortex.

DOI: 10.1103/PhysRevLett.119.038101

The spatial position of an animal can be reliably decoded from the neuronal activity of several cell populations in the hippocampal formation [1–3]. For example, place cells in the hippocampus fire at only a few locations in a spatial environment [4,5], and the position of the animal can be readily read out from single active neurons. Grid cells of the medial entorhinal cortex (MEC) fire at multiple distinct places that are arranged on a hexagonal lattice [6,7]. Although hexagonal patterns are abundant in nature and there exist well-studied physical theories for their emergence, the mechanistic origin of this neuronal grid pattern is still unclear. Initially, it was suggested that they result from continuous attractor dynamics [8,9] or the superposition of plane wave inputs [10] and, based on circuit anatomy, place cells would then result from a superposition of many grid cells [11,12]. More recent experiments, however, reported place cell activity without intact grid cells such that grid cells are not the unique determinants of place field firing [13–17]. Conversely, it would thus be possible that grid fields may arise from place field input as suggested in [18–20]. The biological mechanisms proposed by these latter theories, however, remain hypothetical. In the present Letter, we propose a learning rule for grid cells, based on the individual spike timings of place cells using spike-timing-dependent synaptic plasticity (STDP) [21–23]. The theory thereby predicts that the observed temporal hippocampal firing patterns (phase precession and theta-scale correlations, see below) [24–26] translate the temporal proximity of sequential place field spikes into spatial neighborhood relations, observed in grid-field activity. For our model to work, we only have to assume that the synaptic plasticity rule averages over a sufficiently long time interval.

Model.—We use the classical formulation of pairwise additive STDP [22,27], where the update of a synaptic weight J_n, $n = 1, \ldots, N$ at time t is computed as [22]

$$\frac{d}{dt}J_n = \int_{-\infty}^{\infty} ds W(s) C_n(s) + F(J_n). \tag{1}$$

$C_n(s)$ denotes the time-averaged correlation function between the spike train of the presynaptic neuron n and the postsynaptic neuron, the learning window $W(s)$ describes the update of the synaptic weight as a function of the time difference s between a pair of pre- and postsynaptic action potentials, and the function F implements soft bounds for the weight increase. The dynamics is further constrained such that weights cannot become negative.

To be able to treat Eq. (1) analytically, we use a linear Poisson neuron model; i.e., the mean firing rate of the postsynaptic neuron $E(t) = \mathbf{J} \cdot \mathbf{H}(t)$ results from a weighted sum of hippocampal firing rates $\mathbf{H} = (H_1(t), \ldots, H_N(t))^T$. Under these assumptions, $C_n(s)$ can be approximated for large N [22] as $C_n(s) = \sum_{n'} J_{n'} C_{nn'}(s)$, with

$$C_{nn'}(s) := \int_{-\infty}^{\infty} dt H_n(t) H_{n'}(t - s). \tag{2}$$

Inserting the correlation functions from Eq. (2) into the weight dynamics from Eq. (1) yields

$$\frac{d}{dt}J_n = \sum_{n'} J_{n'} G_{nn'} + F(J_n), \quad G_{nn'} := \int_{-\infty}^{\infty} ds W(s) C_{nn'}(s). \tag{3}$$

Following [28,29], we introduce the quadratic stabilization term $F(J) = F_0 J(K - J)$, $F_0 > 0$ that implements a soft upper bound.

As an input to the postsynaptic neuron, we consider a population of N hippocampal place cells. The firing of these neurons is characterized by a bell-shaped envelope, modulating the spatial path $x_{\mathcal{P}}(t)$ and oscillations in time t [Fig. 1(a)]

$$H_n(t; \mathcal{P}) = a e^{\frac{-[x_{\mathcal{P}}(t) - x_n]^2}{2\sigma^2}} [\cos(\omega t + \phi_n) + 1]/2. \tag{4}$$

The oscillation frequency ω of a neuron is slightly higher than the frequency $\omega_\theta/(2\pi)$ of the theta oscillation (~8 Hz)

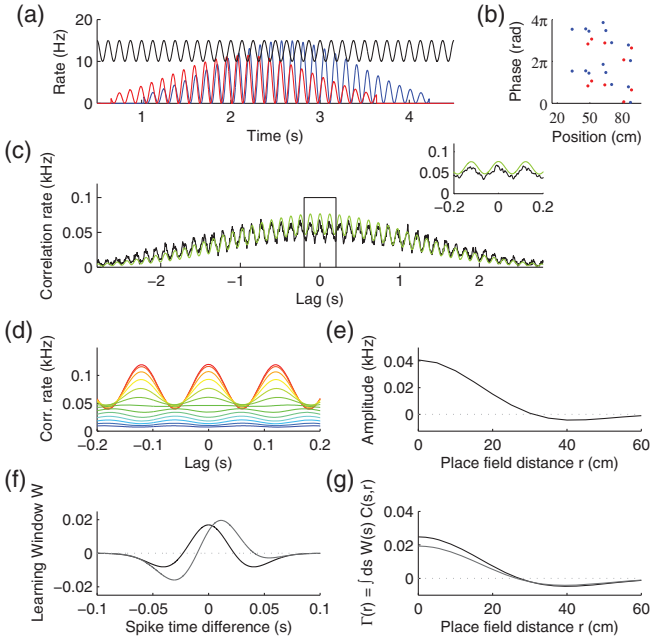


FIG. 1. (a) Poisson model of place field firing for two place cells (red, blue) and slower theta oscillation (black). (b) Phase precession resulting from the path used in (a). (c) Correlation of spike trains for the cells in (a), averaged over 2D trajectories (random walk) (black) and from Eq. (5) (green). Inset magnifies lag 0. (d) Correlation functions for different place field distances (red: autocorrelation, blue: 60 cm). (e) Correlation amplitude as a function of place field distance. (f) Examples of STDP learning windows, see Eq. (9) and below. (g) Γ kernels for correlation function from (c)–(e) and learning windows from (f).

in the local field potential, giving rise to a phenomenon called theta phase precession [Fig. 1(b)]: spikes early in the field come at later phases than spikes late in the field [24]. During the traversal of a place field, phase precession spans a whole theta cycle [30]. Thus, the two frequencies have to relate to each other like $\omega = \omega_\theta + (\pi/R)v$, with R denoting the distance from the place field center, at which, the firing rate has decreased to 10% [i.e., $R = \sigma\sqrt{2\ln(10)}$], and v denoting the running speed, which we fix at 25 cm/s. At each individual entry into the place field, the phase of the cellular oscillation is reset to phase zero with respect to the theta oscillation phase ϕ_θ by fixing $\phi_n = (\phi_\theta - 2\pi)\omega/\omega_\theta$.

To obtain a closed expression for the correlation function $C_{nn'}(s)$ [Fig. 1(c)], the time average in Eq. (2) is performed over all straight paths $\mathcal{P}$ crossing the center of the field overlap. For place fields with identical width R, firing rates a, and at small lags s, we obtain

$$\begin{aligned} C_{nn'}(s) &= \int_{\mathcal{P}} \int_{-\infty}^{\infty} dt H_n(t;\mathcal{P}) H_{n'}(t-s;\mathcal{P}) \\ &= a\frac{\sqrt{\pi}e^{-(r^2+v^2s^2/4\sigma^2)}}{4v/(a\sigma)}\left[1+\frac{1}{2}J_0\left(\frac{\pi r}{R}\right)\cos(\omega s)\right. \\ &\quad \left. + \frac{vsr}{4\sigma^2}J_1\left(\frac{\pi r}{R}\right)\sin(\omega s)\right], \end{aligned} \tag{5}$$

with place field distance $r = |x_n - x_{n'}|$ and $J_{0,1}$ denoting Bessel functions of the first kind (see Supplemental Material [31]).

In contrast to 1D fields, where the distance of place fields is reflected by the lag of the correlation peak (theta compression) [25], correlation functions in 2D are symmetric because of the symmetry of the path; however, the distance of the place field centers is encoded in the amplitude of the correlation peak at lag 0 [Figs. 1(d) and 1(e)].

Weight Dynamics.—Assuming that the putative grid cells receive inputs from a large number $N \gg 1$ of place cells that sufficiently cover the encoded area, we replace the presynaptic index n by the position of the place field center x; i.e., $G_{nn'} \to \Gamma(|x_n - x_{n'}|)$, and thereby translates the learning Eq. (3) to continuous coordinates

$$\frac{d}{dt}J(x) = (\Gamma \circ J)(x) + F_0 J(x)(K - J). \tag{6}$$

Examples of the convolution kernel $\Gamma(|x|)$ for different learning window functions W are depicted in Figs. 1(f) and 1(g). The development of the weights follows the pattern formation principles of a lateral inhibition system [32]. Indeed, the integro-differential equation (IDE) (6) involves nonlocal interactions effectively implemented through the convolution kernel, inducing a strong close-range potentiation and a weaker long-range depression of neighboring synapses, as observed in the typical shape presented in Fig. 1(g). A general window-dependent kernel can be obtained for the correlation function from Eq. (5) as

$$\Gamma(r) = c\frac{\sqrt{\pi}e^{-(r^2/4\sigma^2)}}{4v/(a\sigma)}\left[1 + \alpha J_0\left(\frac{\pi r}{R}\right) + \frac{\beta r}{\sigma}J_1\left(\frac{\pi r}{R}\right)\right] \tag{7}$$

$$\begin{aligned} c[W] &:= a\int_{-\infty}^{\infty}\exp\left[-\frac{v^2s^2}{4\sigma^2}\right]W(s)ds \\ \alpha[W] &:= \frac{a}{2c}\int_{-\infty}^{\infty}\exp\left[-\frac{v^2s^2}{4\sigma^2}\right]\cos(\omega s)W(s)ds \\ \beta[W] &:= \frac{av}{4\sigma c}\int_{-\infty}^{\infty}\exp\left[-\frac{v^2s^2}{4\sigma^2}\right]s\sin(\omega s)W(s)ds, \end{aligned} \tag{8}$$

which can take a Mexican-hat-type shape for qualitatively different learning window functions W due to the symmetry of the correlation function [Figs. 1(f) and 1(g)]. To see this, we can regard Hebbian-like windows to be modeled as the product of a Gaussian and a polynomial of some order m, $W(s) = \exp[-s^2/(2\rho^2\mu^2)]P_m(s/\rho)$. The functionals defined in Eq. (8) then inherit the symmetries from the cross-correlation since all of the odd terms in the polynomial cancel out during integration. Thus, in subsequent numerical investigations, we focus on windows up to second polynomial order

$$W(s) = W_0(2\pi\mu^2\rho^2)^{-\frac{1}{2}}[1 - (s/\rho)^2]e^{-(s^2/2\rho^2\mu^2)}, \tag{9}$$

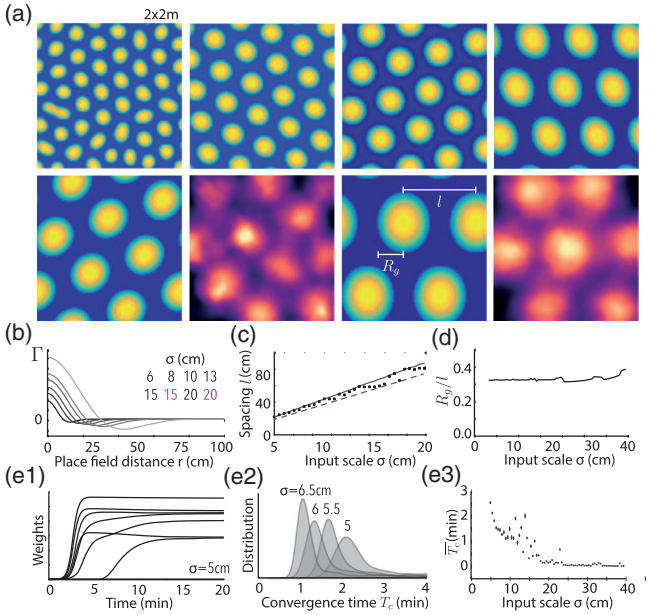


FIG. 2. Weight patterns and input scale. (a) Asymptotically stable weight functions $J(x)$ for place field widths σ, as indicated in (b), together with the respective Γ kernels ($\mu = 1.025$, $\rho = 23$ ms, σ indicated by grey level). Firing rates of spiking simulations are colored in red (for numerical details, see Supplemental Material [31]). (c) Grid spacing l [see (a)] scales linearly with place field scale σ. Dots correspond to weight patterns from numerical solutions. Dashed and solid lines indicate theoretical estimates $2\pi/k_m$ and $2\pi/\tilde{k}_m$, see linear theory at early times and positivity constraint. (d) Estimated ratio of field radius R_g [see (a)] to grid spacing l; $R_g/l = [\sqrt{3}\nu/(2\pi)]^{1/2}$ is independent of σ. (e1) Temporal evolution of randomly sampled weights. (e2) Distribution of weight convergence times T_c for different input scales σ as indicated and the respective means (e3).

whose free parameters ρ and μ determine its zeroes $s_0 = \pm\rho$ and negativity $\int W = W_0(1-\mu^2)$. In Figs. 1(f) and 1(g), we used $\rho = 23$ ms, $\mu = 1.025$ and added a linear term s/ρ to the polynomial to get the asymmetric window (grey lines). The W-dependent functions c, α, and β defining the kernel Γ are given in the Supplemental Material [31].

A numerical evaluation of the learning IDE (6) with periodic boundary conditions reveals that the spatially isotropic kernel Γ can result in hexagonal packing structures [Figs. 2(a) and 2(b)]. Simulations of spiking Poisson neurons confirm these predictions of the meanfield theory [Fig. 2(a)]. As indicated by the kernel function in Eq. (7), the grid spacing only depends on the spatial scale σ of the place fields in the input [Fig. 2(c)].

An experimentally accessible quantity to compare our model results to is the ratio of grid field radius R_g to grid spacing l as indicated in Fig. 2(a). In experiments, the ratio R_g/l for grid cells has been determined to be about 0.3 [7], and for a perfectly hexagonal grid, R_g/l relates to the fraction ν of field size per area as $\nu = \pi R_g^2/[(\sqrt{3}/2)l^2]$. The field fraction ν can be readily accessed from our numerics as the fraction of nonzero synaptic weights and for the used learning window, fits the experimentally obtained R_g/l [Fig. 2(d)] for all choices of σ. Finally, learning converges faster for large spacing (large σ), consistent with a larger amplitude of Γ [Eq. (7) and Figs. 2(b) and 2(e)].

Linear theory at early times.—Some analytical understanding of the weight dynamics from Eq. (6) can be gained from a neural field theory approach [32–36]. In this framework, we can neglect the effect of the nonlinearities at early times and focus only on the convolution term $\Gamma * J$. The emerging dynamics can be readily understood by looking at the evolution of the weights in Fourier space $\partial_t \hat{J}(k) = \hat{\Gamma}(k)\hat{J}(k)$, which makes evident that the wave number k_m maximizing the kernel Fourier transform $\hat{\Gamma}$ (see Supplemental Material [31]) will exponentially overgrow all other modes [if $\hat{\Gamma}(k_m) > 0$], thus setting the initial periodicity of the pattern.

In a finite region of the parameter space (μ, ρ) of the learning window, the bimodal (Mexican hat) shape of Γ ensures the existence of a Turing instability, i.e., a transition to a single maximum of $\hat{\Gamma}$ [with $\hat{\Gamma}(k_m) > 0$] at a nonzero $k_m > 0$ [Figs. 3(a) and 3(c)]. Similar to previous work on pattern formation in lateral inhibition systems (e.g., [37]), the permitted parameter region (k_m exists and is positive) gives rise to stripelike and hexagonal patterns [Fig. 3(d)]. In the Supplemental Material [31], we also provide a complementary description of the pattern formation process, based on an approximation of Eq. (6) by a partial differential equation (Swift-Hohenberg equation [38]), which corroborates the results from the linear theory.

To connect the resulting patterns to other feed-forward models of grid field formation, we parameterize Γ by the shape factor r_0/r_m [Fig. 3(c)], which is the fraction between the zero and the minimum of Γ. The shape factor r_0/r_m reduces the two-parameter learning window to a single qualitatively descriptive parameter, which can be used to describe the bimodal kernel Γ independently of the hypothesized biological mechanism. If r_0/r_m is large ($\sim$0.8), Γ shows only little negativity, and the emerging pattern is stripelike [Figs. 3(c) and 3(d)]; if r_0/r_m is small, Γ exhibits strong negativity, the firing fields become dispersed, and the pattern looses hexagonality. Hexagonal patterns arise for r_0/r_m roughly between 0.65 and 0.75 [Fig. 3(d)]. In this region, the shape factor virtually completely determines the geometrical properties of the steady state [Figs. 3(e) and 3(f)]. Values of r_0/r_m that give rise to hexagonal grids can also be identified via the ratio of field width per grid spacing R_g/l. According to our theory, the experimentally observed value 0.3 [7] is achieved with a shape factor of about $r_0/r_m = 0.7$ [Fig. 3(f)]. For higher values of r_0/r_m, R_g/l increases to a point where a periodic pattern cannot further dissociate into disjoint fields, and the stable pattern becomes stripelike. For lower r_0/r_m, R_g/l decreases, and, at some point, the small fields no longer repel each other strongly enough to produce a symmetrical arrangement.

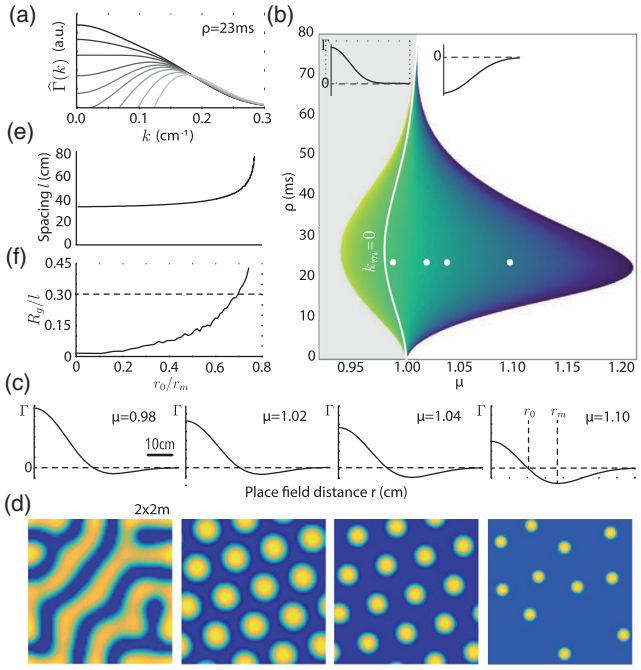


FIG. 3. Permitted learning windows. (a) Fourier transforms of Γ for $\mu = 0.95$ (black) to 1.20 (light grey). (b) Region of structure formation in (μ, ρ) space (for $\sigma = 10$ cm). Nontrivial patterns appear to the right of the white solid line ($k_m = 0$), where the selected wavelength is positive. Colored region indicates bimodal Γ. The color encodes the shape factor r_0/r_m (0: blue, 1: yellow). Grey and white areas correspond to regions where Γ is all positive or all negative. (c) Four examples of Γ kernels, corresponding to the white dots from (b) in the same order. (d) Weight patterns for kernels from (c). (e) Predicted spacing $2\pi/k_m$ as a function of r_0/r_m, obtained for all combinations of μ and ρ from the region right of the solid line in (a). (f) Estimated ratio of field size to grid spacing as a function of r_0/r_m (from simulated patterns). Dashed line indicates experimental value [7].

Positivity constraint.—The grid spacing l predicted by the linear theory, however, consistently underestimates the spacing derived from the numerical solution of the mean field dynamics [Fig. 2(c)]. The reason for this error is that, after the initial growth phase, the synaptic weights are influenced by the nonlinearities, most importantly the constraint that they cannot become negative.

The impact of this positivity constraint can be intuitively understood if we interpret the convolution $\Gamma \circ J$ as an operation that detects the best overlap of an oscillatory pattern $J \propto \cos(kx)$ with a given kernel Γ. However, after the lowest weights reach zero, they stop contributing to the convolution, and a slightly lower wave number $\tilde{k}_m$ maximizing

$$\tilde{\Gamma}(k) := \int_\Omega dx\, \Gamma(x) \cos(kx) \Theta[\cos(kx) - \cos(|k|R_g)] \quad (10)$$

will be favored as the fastest growing mode (Θ denoting the Heaviside function). Similarly, a particular field size R_g maximizing $\tilde{\Gamma}$ will be selected. In the experimentally relevant case $|k|R_g = 2\pi R_g/l = 2\pi \times 0.3$, numerical maximization of Eq. (10) yielded the predicted wave number $\tilde{k}_m$ [solid line in Fig. 2(c)], which excellently agrees with the numerical solutions of the meanfield dynamics.

Conclusion.—For a large variety of STDP windows, the spike-timing correlations of 2D place cells can account for a feed-forward learning of hexagonal grid patterns. Synaptic plasticity thereby averages over running trajectories of tens of minutes, hence, translating the temporal correlations into a dense code for space. Our model thus predicts that grid cells are generated in the output structures of the hippocampus, e.g., the deep layers of the medial entorhinal cortex [39] or the parasubiculum [40]. While our linear theory provides a good prediction of grid spacing, as well as for conditions that permit structure formation, determining the boundary between hexagonal and stripe-like patterns is less straightforward and has to take into account the nonlinearities. The standard approach, nonlinear bifurcation analysis [37,41,42], is difficult because of the strong nonlinearity introduced via the positivity constraint, which strongly influences the selection of the final pattern. Despite this drawback, our model provides a universal framework in that it encompasses current models of grid field formation that can be mapped to convolutions with Mexican-hat-type kernels that give rise to a Turing instability.

The authors are grateful to Andreas Herz and Anton Sirota for discussions and Martin Stemmler for comments on the manuscript. This work was funded by the German Research Foundation (DFG), Grant No. LE2250/5-1.

[1] K. D. Harris, J. Csicsvari, H. Hirase, G. Dragoi, and G. Buzsaki, Nature (London) **424**, 552 (2003).
[2] A. Mathis, A. V. Herz, and M. Stemmler, Neural Comput. **24**, 2280 (2012).
[3] A. Kammerer and C. Leibold, PLoS Comput. Biol. **10**, e1003986 (2014).
[4] J. O'Keefe, Exp. Neurol. **51**, 78 (1976).
[5] R. U. Muller, J. L. Kubie, and J. B. Ranck, J. Neurosci. **7**, 1935 (1987).
[6] M. Fyhn, S. Molden, M. P. Witter, E. I. Moser, and M. B. Moser, Science **305**, 1258 (2004).
[7] T. Hafting, M. Fyhn, S. Molden, M. B. Moser, and E. I. Moser, Nature (London) **436**, 801 (2005).
[8] M. C. Fuhs and D. S. Touretzky, J. Neurosci. **26**, 4266 (2006).
[9] Y. Burak and I. R. Fiete, PLoS Comput. Biol. **5**, e1000291 (2009).
[10] N. Burgess, Hippocampus **18**, 1157 (2008).
[11] B. L. McNaughton, F. P. Battaglia, O. Jensen, E. I. Moser, and M.-B. Moser, Nat. Rev. Neurosci. **7**, 663 (2006).
[12] E. I. Moser and M. B. Moser, Hippocampus **18**, 1142 (2008).

[13] J. Koenig, A. N. Linder, J. K. Leutgeb, and S. Leutgeb, Science **332**, 592 (2011).
[14] M. P. Brandon, A. R. Bogaard, C. P. Libby, M. A. Connerney, K. Gupta, and M. E. Hasselmo, Science **332**, 595 (2011).
[15] T. J. Wills, F. Cacucci, N. Burgess, and J. O'Keefe, Science **328**, 1573 (2010).
[16] J. B. Hales, M. I. Schlesiger, J. K. Leutgeb, L. R. Squire, S. Leutgeb, and R. E. Clark, Cell Rep. **9**, 893 (2014).
[17] M. I. Schlesiger, C. C. Cannova, B. L. Boublil, J. B. Hales, E. A. Mankin, M. P. Brandon, J. K. Leutgeb, C. Leibold, and S. Leutgeb, Nat. Neurosci. **18**, 1123 (2015).
[18] E. Kropff and A. Treves, Hippocampus **18**, 1256 (2008).
[19] Y. Dordek, D. Soudry, R. Meir, and D. Derdikman, eLife **5**, e10094 (2016).
[20] A. Stepanyuk, Biol. Inspir. Cogn. Arc. **13**, 48 (2015).
[21] H. Markram, J. Lubke, M. Frotscher, and B. Sakmann, Science **275**, 213 (1997).
[22] R. Kempter, W. Gerstner, and J. L. van Hemmen, Phys. Rev. E **59**, 4498 (1999).
[23] G. Q. Bi and M. M. Poo, J. Neurosci. **18**, 10464 (1998).
[24] J. O'Keefe and M. L. Recce, Hippocampus **3**, 317 (1993).
[25] G. Dragoi and G. Buzsaki, Neuron **50**, 145 (2006).
[26] T. Feng, D. Silva, and D. J. Foster, J. Neurosci. **35**, 4890 (2015).
[27] W. Gerstner, R. Kempter, J. L. van Hemmen, and H. Wagner, Nature (London) **383**, 76 (1996).
[28] M. C. van Rossum, G. Q. Bi, and G. G. Turrigiano, J. Neurosci. **20**, 8812 (2000).
[29] W. M. Kistler and J. L. van Hemmen, Neural Comput. **12**, 385 (2000).
[30] A. P. Maurer, S. L. Cowen, S. N. Burke, C. A. Barnes, and B. L. McNaughton, Hippocampus **16**, 785 (2006).
[31] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevLett.119.038101 for derivation.
[32] N. V. Swindale, Proc. R. Soc. B **208**, 243 (1980).
[33] R. Linsker, Proc. Natl. Acad. Sci. U.S.A. **83**, 7508 (1986).
[34] B. Ermentrout, J. Campbell, and G. Oster, Veliger **28**, 369 (1986).
[35] K. D. Miller, J. B. Keller, and M. P. Stryker, Science **245**, 605 (1989).
[36] H. Shouval, N. Intrator, C. C. Law, and L. N. Cooper, Neural Comput. **8**, 1021 (1996).
[37] G. B. Ermentrout and J. D. Cowan, Biol. Cybern. **34**, 137 (1979).
[38] J. Swift and P. C. Hohenberg, Phys. Rev. A **15**, 319 (1977).
[39] M. P. Witter, Hippocampus **3 S1**, 33 (1993).
[40] T. van Groen and J. M. Wyss, Brain Res. **518**, 227 (1990).
[41] G. B. Ermentrout and J. D. Cowan, SIAM J. Appl. Math. **38**, 1 (1980).
[42] G. B. Ermentrout and J. D. Cowan, SIAM J. Appl. Math. **39**, 323 (1980).

Supplementary Material: Hippocampal Spike-Timing Correlations Lead to Hexagonal Grid Fields
by Mauro M Monsalve-Mercado, Christian Leibold

Notes on the derivation of the Correlation function

The activity of a homogeneous population of place fields, sharing equal spatial scale, cellular oscillation frequency, and peak firing rates is modeled according to eq. (4) of the main text as

$$H_n(t;\mathcal{P}) = a\,\exp\left(\frac{-(x_\mathcal{P}(t)-x_n)^2}{2\sigma^2}\right)\left[\frac{\cos(\omega\,t+\phi_n)+1}{2}\right], \tag{1}$$

where the cellular oscillation phase ϕ_n is locked to the phase ϕ_θ of the theta oscillation of the field potential by setting the entrance phase of the place field $\phi_n = (\phi_\theta - 2\pi)\,\omega_n/\omega_\theta$. The frequency of the place cells ω_n is faster than the frequency ω_θ of the field potential resulting in the phenomenon of phase precession.

We then proceed to derive an analytical expression for the correlation function

$$C_{nn'}(s) \quad = \quad \int_\mathcal{P} C_{nn'}(s,\mathcal{P}) = \int_\mathcal{P}\int_{-\infty}^{\infty} \mathrm{d}t\, H_n(t;\mathcal{P})\, H_{n'}(t-s;\mathcal{P}) \tag{2}$$

that would remain valid for a noisy two dimensional random walk with sufficient covering of the place fields' overlap region.

To this end, we start by computing the correlation on a straight path traversing the place fields, $x_\mathcal{P}(t) := x_0 + v\,n\,t$, where v is the speed, x_0 a reference point, and n a two dimensional unit vector specifying the running direction. Considering typical values for all physiological parameters involved, we can approximate the result as

$$C_{nn'}(s,\mathcal{P}) = \frac{\sqrt{\pi}}{4}\frac{\sigma a^2}{v}\exp\left\{-\frac{c_\mathcal{P}^2}{\sigma^2} - \frac{v^2 s^2}{4\sigma^2} - \frac{r^2}{4\sigma^2} - \frac{vsr\cos(\gamma_\mathcal{P})}{2\sigma^2}\right\}\left\{1+\frac{1}{2}\cos(\omega\tau + \delta\phi(\phi_\theta, c_\mathcal{P}, \gamma_\mathcal{P}))\right\}, \tag{3}$$

where $r := |x_n - x_{n'}|$ is the distance between the place fields' centers, $c_\mathcal{P}$ is the perpendicular distance between the center of the overlap $\frac{1}{2}(x_n + x_{n'})$ and the straight path, and $\gamma_\mathcal{P}$ is the angle made between the line joining the place fields' centers and the path (see Supplementary Figure 1). The phase difference $\delta\phi := \phi_n - \phi_{n'}$ is computed from that phases at which path enters each of the place fields. As specified above, these entrance phases are obtained from the instantaneous theta phase at the moment the animals enters the specific field.

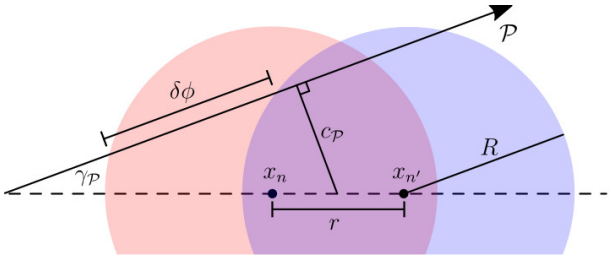


Figure 1: Geometry of two place fields (red,blue). Parameters are explained in the text.

The approximated 2-d correlation function finally results from the aggregated contribution of all possible straight paths crossing the overlapping area of the place fields. The approximation is valid for a sufficiently long run, such that a representative sample of orientations and starting points of the paths, as well as the onset theta phases is taken into account. As it turns out, we can restrict ourselves to paths crossing the center of the overlap ($c_\mathcal{P} = 0$) in all directions, since the contribution coming from different path displacements is negligible. After performing an integration over theta phases and orientations ($\gamma_\mathcal{P}$), we can further reduce the result for small time lags s, since we are only interested in the interaction with learning windows with limited range. The final expression

$$C_{nn'}(s) = a\,\frac{\sqrt{\pi}\,\mathrm{e}^{-\frac{r^2+v^2\,s^2}{4\sigma^2}}}{4v/(a\sigma)}\left(1+\frac{1}{2}J_0\left(\frac{\pi r}{R}\right)\cos(\omega s) + \frac{vsr}{4\sigma^2}J_1\left(\frac{\pi r}{R}\right)\sin(\omega s)\right) \tag{4}$$

uses the relation $R = \pi v/(\omega - \omega_\theta)$ for the radius of the field. $J_{0,1}$ denote Bessel functions of the first kind.

The previous mean field approximation can be seen in comparison with a spiking simulation in Fig. 1C. There, a long 20 min random walk is simulated on which spikes are drawn from two place fields using an inhomogeneous Poisson process.

Numerical Methods

Meanfield simulations
To numerically evaluate the dynamical evolution of the weights according to the learning equation (6), we simulated a discrete analogue. For this, we arranged the place field centers in a equally spaced two-dimensional square grid covering the box. The discrete convolution was performed by direct matrix multiplication with a matrix whose entries evaluate the kernel for the euclidean distances between all pairs of place field centers with periodic boundary conditions. Temporal integration was done using the Euler-forward method with time discretized in bins of 10 ms.

Fig. 2A shows six examples out of fifty simulated weight patterns exhibiting stable hexagonality. The scale σ of place field inputs changes in a range from 5 to 30 cm in 0.5 cm steps. The learning window was kept constant. Spacings of hexagonal grids were estimated from the autocorrelograms by computing the radial power and selecting the maximum at the smallest non-zero wave number (dots in Fig. 2C).

To explore the effect of the learning window, we computed the effective kernel for all combinations in a range of ρ between 1 and 80 ms, and μ between 0.95 and 1.20, for 100 values each. We identified regions of bimodality and computed all Fourier transforms and corresponding maxima (Fig. 3A,B,E).

Default values for all meanfield simulations were as follows.

Parameter	Value
Box size	2×2 m
Number of place cells	100×100
Simulation time	30 min
Speed	25 cm/s
Peak firing rate a	30 Hz
Theta frequency ω_θ	$2\pi \times 8$ Hz
K	0
F_0	0.0011 Hz
Window scaling W_0	0.0011 Hz
Window negativity μ	1.025
Window zero ρ	23 ms
Place field size σ	10 cm

Spiking simulations
We simulated place field activity as 2500 independent inhomogeneous Poisson processes with densities

$$\lambda_n(t) \propto \exp\left[-\frac{(vt - c_n)^2}{2\sigma^2}\right] \left[1 + \cos\left(\omega_c \left(t - c_n/v\right)\right)\right]^4,$$

were v denotes running speed, t is time, c_n the place field center, σ specifies place field width, and $\omega_c = 2\pi/T_c$ is the oscillation frequency of the individual neurons. Place field centers were equally spaced across the 2×2 m square arena. In all simulations we assumed a constant speed $v = 25$ cm/s. The oscillation period T_c of a cell was width dependent to ensure a width independent phase range, $1/T_c = 1/T_\theta + \frac{v}{2R}$, with a theta period of $T_\theta = 1/8$ s and the field radius $R = \sigma\sqrt{2\ln(10)}$.

The neuronal spiking and the running trajectory of the rat in a 2×2 m square arena was simulated with a time resolution of 5 ms. In each time step the running direction was changed by random angle drawn from a Gaussian distribution with standard deviation of 10°. Angles that led to wall collision were not permitted.

The entorhinal activity was simulated by a Poisson process derived from a rate density $E = \sum_{n=1}^{2500} J_n H_n$, where H_n was 0 or 1, depending on whether hippocampal cell n was spiking in the respective time bin or not.

The synaptic weights were initialized at time 0 from a Gaussian distribution with mean 0.025 and a standard deviation 0.001. Pairwise STDP according to Eq. (1) was averaged over 2.5 s before the weight was updated. As an STDP learning rate we used $W_0 = 0.0008$ per spike pair. The non-linear saturation was parametrized by $K = 0$ and $F_0 = 0.01$ Hz. Negative weights were generally always clipped to zero.

Approximate equivalence to the Swift-Hohenberg partial differential equation

The integro-differential equation (6), which determines the structure formation process of the synaptic weight vectors does not make explicit the non-negativity constraint on the weights, $J(x) \geq 0$. Mathematically this constraint can be dealt with by introducing an exponential recasting $J(x) = \exp[\psi(x)]$, and the resulting IDE for ψ then reads

$$\dot{\psi} = J^{-1}\,\dot{J} = \int \mathrm{d}\xi\,\Gamma(\xi)\,[e^{\psi(\xi+x)} - e^{\psi(x)}]\ .$$

If we expand ψ up to fourth order in $\xi = (\xi^{(1)}, \xi^{(2)})^T$ (assuming limited spatial extent of Γ) and the exponential function up to linear order in ψ (assuming small amplitudes), we obtain

$$\begin{aligned}\dot{\psi}(x) &= \int \mathrm{d}\xi\,\Gamma(\xi) + \frac{1}{2!}\int \mathrm{d}\xi\,\Gamma(\xi)\sum_{i,j}\xi^{(i)}\,\xi^{(j)}\partial_{ij}\psi(x) \\ &\quad + \frac{1}{4!}\int \mathrm{d}\xi\,\Gamma(\xi)\sum_{i,j,k,l}\xi^{(i)}\,\xi^{(j)}\,\xi^{(k)}\,\xi^{(l)}\partial_{ijkl}\psi(x)\end{aligned}$$

because odd powers in ξ vanish owing to the radial symmetry of Γ.

Defining the kernel moments

$$\Gamma_{mn} := \frac{1}{m!n!}\int_\Omega \Gamma\left(\sqrt{(\xi^{(1)})^2 + (\xi^{(2)})^2}\right)\,(\xi^{(1)})^n\,(\xi^{(2)})^m\ ,$$

the learning equation reduces to

$$\dot{\psi} = \Gamma_{00} + \Gamma_{20}\Delta\psi + \Gamma_{40}(\partial_1^4 + \partial_2^4)\psi + \Gamma_{22}\partial_1^2\partial_2^2\psi$$

since due to the radial symmetry, $\Gamma_{20} = \Gamma_{02}$, $\Gamma_{40} = \Gamma_{04}$. Radial symmetry of Γ also yields $\Gamma_{22} = 2\,\Gamma_{40}$ and thus we end up at the linear dynamics

$$\begin{aligned}\dot{\psi} &= \Gamma_{00} + \Gamma_{20}\Delta\psi + \Gamma_{40}\Delta^2\psi \\ &= \Gamma_{00} - \Gamma_{40}\left[\frac{\Gamma_{20}^2}{4\Gamma_{40}} - \left(\Delta + \frac{\Gamma_{20}}{2\Gamma_{40}}\right)^2\right]\psi\ .\end{aligned}$$

Introducing the abbreviations $\varepsilon = \Gamma_{02}^2/(4\Gamma_{04}^2)$ and $k_c^2 = \Gamma_{02}/(2\Gamma_{04})$, we rewrite the previous equation as

$$\frac{-1}{\Gamma_{40}}\,\dot{\psi} = \frac{-\Gamma_{00}}{\Gamma_{40}} + [\varepsilon - (k_c^2 + \Delta)^2]\psi\ ,$$

which is known from the Swift-Hohenberg theory of pattern formation used to model Rayleigh-Bénard convection [37]. This theory also includes stripe and hexagonal pattern formation with the necessary conditions that both ε and $-1/\Gamma_{04}$ be positive. The resulting regions of structure formation correspond well to that of the linear theory depicted in Figure 3, in particular the white solid line separating the homogeneous mode from the regime of structure formation matches $\varepsilon = 0$. The wave number k_c sets the grid spacing and also approximates well the simulated spacings from Figure 2C (see Supplementary Figure 2).

For our model used in this Letter, we find the following explicit expression for the kernel moments,

$$\Gamma_{mn} = \frac{a\pi^{3/2}c\sigma^{m+n+3}}{v\,(n/2)!(m/2)!}\,(1 + \alpha\,A_{mn} + \beta\,B_{mn})\,,$$

with constants $A_{mn} := \mathrm{e}^{-\zeta^2}\,{}_1F_1\left(-\frac{m+n}{2}, 1, \zeta^2\right)$, $B_{mn} := (m+n+2)\,\zeta\,\mathrm{e}^{-\zeta^2}\,{}_1F_1\left(-\frac{m+n}{2}, 2, \zeta^2\right)$, and $\zeta = \pi\sigma/R = \pi/\sqrt{2\,\ln(10)} \approx 1.46$. The confluent hypergeometric functions of the first kind ${}_1F_1$ reduce to polynomials of order $m+n$ on ζ.

The constants c, α, β depend mostly on the learning window (see Section Mathematical Details below).

Mathematical Details

The convolution kernel Γ from eq. (7) involves three functionals from eqs. (8), which, for the particular choice of learning window from eq. (9), read

$$c[W] = \sqrt{2}\sigma a\,W_0\frac{2\sigma^2(1-\mu^2) + v^2\rho^2\mu^2}{(2\sigma^2 + v^2\rho^2\mu^2)^{3/2}}$$

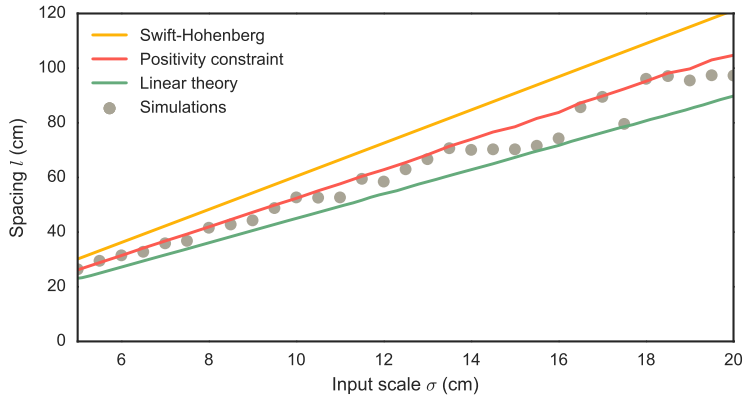


Figure 2: Grid spacing l as a function of σ. Grey dots as well as the red and green lines are copied from Figure 2C of the main text. The orange line estimates the grid spacings using k_c from the Swift-Hohenberg-type partial differential equation given in the supplementary material above.

$$\begin{aligned}\alpha[W] &= Q\left(1+\frac{4\sigma^4\omega^2\rho^2\mu^4}{(2\sigma^2(1-\mu^2)+v^2\rho^2\mu^2)(2\sigma^2+v^2\rho^2\mu^2)}\right)\\ \beta[W] &= \frac{Q\,v\sigma\rho^2\omega\mu^2}{(2\sigma^2+v^2\rho^2\mu^2)^2}\left(2\sigma^2(1-3\mu^2)+v^2\rho^2\mu^2+\frac{4\sigma^4\mu^4(\omega^2\rho^2-2)}{2\sigma^2(1-\mu^2)+v^2\rho^2\mu^2}\right)\end{aligned}$$

with $Q=\frac{1}{2}\exp\left(-\frac{\omega^2\sigma^2\rho^2\mu^2}{2\sigma^2+v^2\rho^2\mu^2}\right)$.

The above quantities determine the Fourier transform of the kernel (below $\mathcal{O}(k)$)

$$\hat{\Gamma}(k)=c\frac{\sqrt{\pi}\sigma^2\mathrm{e}^{-\sigma^2k^2}}{2v/(a\sigma)}\left(1+\alpha\,\mathrm{e}^{-(\frac{\sigma\pi}{R})^2}\,I_0\left(2\pi\frac{\sigma^2k}{R}\right)\right)$$

with I_0 denoting the modified Bessel function of the first kind.

Since $\sigma/R=1/\sqrt{2\ln(10)}$ is a constant, the maximum of $\hat{\Gamma}(k)$, and hence structure formation, will be mostly determined by α. Furthermore, since for physiological parameters and $\mu>1$, α almost entirely depends on properties of the learning window W, the formation of hexagonal patterns crucially depends on W and only little on the input correlations. This explains, why not all structures receiving place field activity may develop grid cells, but only those that have the fitting STDP rule.

PART III

DISCUSSION

A reality check

How well can CANs path-integrate?

Path integration is regarded as a landmark success of continuous attractor models. However, how do these models fare when exploring an environment? Burak and Fiete (2006) put to the test the original Fuhs and Touretzky (2006) attractor model to evaluate the performance of its path integration mechanism. They characterised the deviations of the activity on a real open field trajectory with respect to changes in the neural sheet. They found the tracking in the neural sheet of the animal's movement to accumulate errors quickly, destroying the hexagonal pattern in space for typically recorded running times and lengths.

In further work, Burak and Fiete (2009) modified slightly the original model to show that it is possible to obtain accurate path integration for typical behavioural open field recordings. Their most important modification is to consider an activity-based (or firing rate) model instead of the original voltage-based model used by Fuhs and Touretzky (2006). The inclusion of the velocity coupling inside the rectifying non-linearity greatly improves the accuracy of path integration. This improvement is due to the intrinsic difference on how inputs are integrated in the dynamics (Ermentrout and Terman 2010). In the original model a slow time constant dominates the dynamics, meaning that rapid changes in the input are not properly captured and greatly contribute to integration errors. In comparison, the firing rate model [7] actually integrates over a low-pass-filtered version of the velocity coupling input, removing in this way the rapid changes causing error accumulation, and otherwise has been shown to be mathematically equivalent to the voltage-based alternative (Miller and Fumarola 2012).

In addition, their connectivity profile is modelled after a difference of Gaussians (a standard Mexican-hat profile) with an offset from the cell's centre. How well the attractor networks perform path integration is illustrated in Figure 21 for periodic and aperiodic-damped boundary conditions.

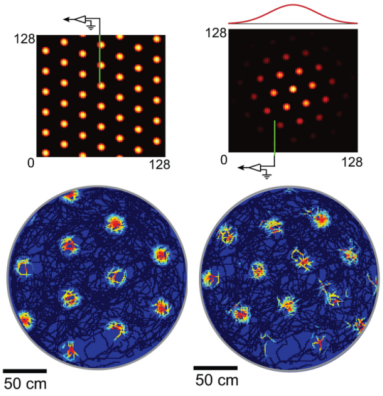


Figure 21: Path integration on CANs *is shown to be more accurate with periodic boundary conditions (left) when compared to aperiodic ones modulated by a decay close to the edges (right). Modified with permission from Burak and Fiete 2009.*

[7] The output of this model might be better interpreted as a low-pass-filtered version of the actual firing rate of the cell.

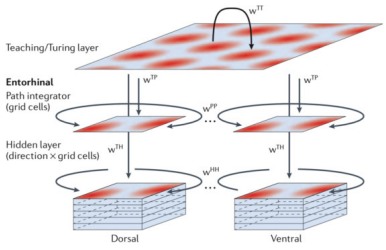


Figure 22: Learning a toroidal topology. *A transient population of hard-wired grid cells act as a teaching layer for topographically organised modules of downstream neurons. A toroidal connectivity configuration subsequently arises via Hebbian learning. Modified with permission from McNaughton et al. 2006.*

One evident shortcoming of the models are the strong constraints imposed on the actual or effective connectivity among stellate cells in dMEC. The mechanisms by which such a specific connectivity profile is achieved must agree with the known anatomical and physiological characteristics of the entorhinal-hippocampal loop. Fuhs and Touretzky (2006) suggested that, similar to the visual system, randomly oriented wave packets of activity would lead to such a configuration via plastic changes. As in neocortex and the retina, waves of activity have been reported in postnatal slices of EC, and moreover topographically organized synchronous activity appears for a transient period of time prior to eye opening in rodent pups (Dawitz et al. 2017; Egorov and Draguhn 2013). These findings also lend some support to the proposal by McNaughton et al. (2006) that the effective connectivity in MEC has a toroidal topology. Illustrated in Figure 22, such a configuration would be learned through Hebbian plasticity between topographically organized modules of cells receiving direct instruction from a teaching layer. The idea borrows from other neural systems exhibiting transient patterns of connectivity or reallocation of cell populations during early development. A specific population forming the teaching layer would presumably display a hard-wired hexagonal pattern of activity, and exploration of all phases by noise-induced drift would teach the desired toroidal topology to the downstream population.

Another way to exploit a periodic connectivity is to consider a single bump of activity that repeats itself when traversing the environment. Guanella, Kiper, and Verschure (2007) described a model where the connectivity resembles a twisted torus to account for hexagonality. Even though the arrangement has appealing properties (for example requires few neurons for spatial encoding), the pattern of activity would necessarily be perfectly regular, in contrast to actual recorded data. In general, periodic connectivity assumptions can not account for irregularities and defects in the observed patterns, such as penta/hepta pairs and lattice distortions. As mentioned before, it is possible to obtain attractor states without periodic connectivity at the cost of less accurate path integration, and more importantly, over-excitation close to the boundaries predicts a large population of cells steadily firing all over the environment.

A second caveat concerns more realistic assumptions of MEC cell populations. The effective Mexican-hat connectivity can be better interpreted as an approximation of the interaction of excitatory and inhibitory populations, reflecting a more realistic view of neurons in MEC. Indeed, layer II stellate cells are excitatory and exhibit no direct connections among them, but rather interact strongly only through inhibitory interneurons, with several stellate cells projecting to one

interneuron which in turn projects to many stellate cells (Couey et al. 2013; Pastoll et al. 2013). In a more realistic model, Widloski and Fiete (2014) investigate the sudden appearance of attractor states in an excitatory-inhibitory (E-I) network. The connectivity between E-I and I-I populations is learned through spatial exploration of the environment using information from cells tuned to spatial and self-motion cues. The learning is gradually implemented with asymmetric spike-timing dependent plasticity rules acting on stochastic spiking activity to reflect more sensible learning times and dynamics. The quality of path integration for a mature network is shown in Figure 23.

Grid cells also exhibit a marked temporal structure of spiking activity, including strong theta modulation and phase precession (or locking) phenomena. To reconcile these observations with the attractor idea, Navratilova et al. (2012) considered a two-network model with realistic current dynamics. One network, the conjunctive cell layer, receives strong theta input and self-motion information (head-direction and speed signals), and it is tasked with translating the pattern of activity for path integration. The conjunctive layer is coupled to a grid cell layer implementing simple after-spike dynamics. Both networks follow leaky-integrate-and-fire dynamics with a variety of physiological properties of MEC stellate cells including after-depolarisation and medium after-hyperpolarisation. In addition, conductance changes were taken into account for AMPA and NMDA synapses whose time constants control not only temporal phenomena but also predict the increasing grid spacing in the dorso-ventral axis. However, the requirement that stellate cells communicate through excitatory synapses is not supported experimentally. To tackle this, Pastoll et al. (2013) modelled a twisted torus attractor network with excitatory and inhibitory populations and AMPA and NMDA conductances to show that the model is still compatible with theta modulation and additional theta-nested gamma oscillations. In their model, feedback inhibition is key for the coexistence of temporal and spatial structures.

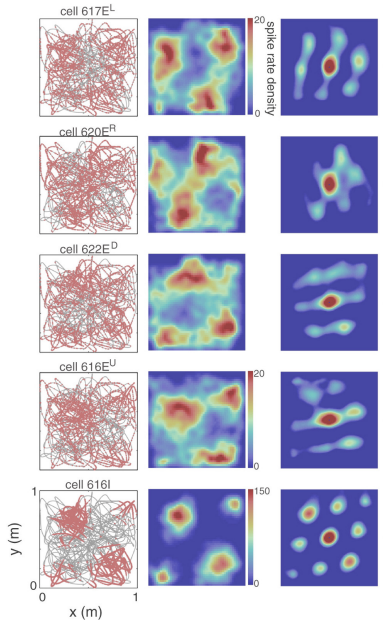


Figure 23: Integration on a E-I model. *Population weights are learned via STDP through teaching from existing spatially modulated cells. The pattern is later maintained by the MEC population alone. Modified with permission from Widloski and Fiete 2014.*

What else can we learn from OIMs?

The temporal structure of grid cell firing can be more suitably addressed in the framework of oscillatory interference models. Strong theta modulation and phase precession are key components on top of which the model is built upon. By construction the model excels in simplicity and interpretability of its parameters, while at the same time addressing the most relevant temporal and spatial properties of grid cell activity. However, the model has met with some criticism

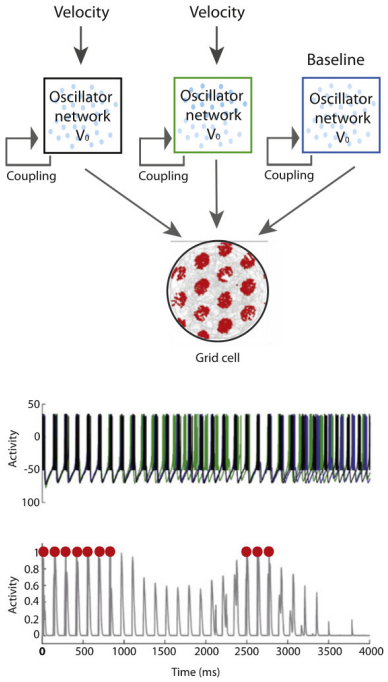


Figure 24: Coupled VCO model. *Top: Several oscillator networks transform velocity input into a phase code in a preferred direction (periodic bands of activity). One of the networks is not modulated by velocity and sets the baseline frequency. All networks are internally coupled but are otherwise independent from each other. Bottom: Spiking activity from three different networks (black, blue, and green) going in and out of phase. When they are close to coincidence the summed activity drives the grid cell to fire (red dots in bottom trace). Modified with permission from Giocomo, Moser, and Moser 2011.*

regarding feasible biological implementations of its rather abstract description.

One early critique concerns difficulties with different dendritic membrane potential oscillators coexisting together for meaningful lapses of time (Remme, Lengyel, and Gutkin 2009). Remme, Lengyel, and Gutkin (2010) showed that independent dendritic and somatic oscillators in a realistic biophysical model tend to synchronize (phase-lock) in time-scales below the second. The grid pattern would then disappear unless a phase resetting mechanism (sensory cues) is in place to correct path integration in the order of hundreds of milliseconds. The other VCO implementation, namely external spiking input, would predict cells with stripe-like firing patterns, which have not been (convincingly) recorded anywhere in the hippocampal formation (but see Krupic, Burgess, and O'Keefe 2012).

The second main critique is concerned with noise. In contrast to perfect sinusoidal oscillations, in vitro slice recordings of entorhinal cells show that subthreshold membrane potential oscillations have a substantial degree of noise and a relatively wide frequency spectrum (in the theta band). Welinder, Burak, and Fiete (2008) pointed out that the phase of theta tends to be very irregular, with its temporal cross-correlation decaying after only a few cycles. Together with the additional effect of neuronal noise, both deviations from perfect oscillators drift the activity from its expected location in just a few seconds, shifting the pattern to the detriment of hexagonality (Zilli et al. 2009).

Zilli and Hasselmo (2010) suggest the previous critiques can be dealt with by moving away from single neuron models. They demonstrate how several independent networks are able to implement robust VCOs under realistic levels of noise (see Figure 24). All Neurons in a given network are connected to each other (total coupling), thus they tend to synchronize their activity in response to common input while maintaining stable firing even for high levels of noise. All networks, however, act independently and do not interact among themselves. One of the networks does not receive sensory input and plays the role of the baseline oscillator network. All other networks receive velocity input and thus implement different VCOs with particular preferred directions. As before, networks differing by sixty degrees add up in the target cell making the hexagonal pattern.

Other modelling attempts have a different take on the oscillatory interference mechanism. Hasselmo (2008) showed that the model can be implemented by a population of neurons exhibiting persisting firing. Some Pyramidal cells in layers III and V of entorhinal cortex respond to a stimulus by a continuous spiking activity after the stimulus is turned off (Egorov et al. 2002). Appropriate modulation

by head direction cells and phase resetting mechanisms enable the population to combine into a phase encoding system of spatial information. However, given the variability in spiking frequency, the model suffers the same sensitivity to noise that affects earlier formulations of the model. Another interesting approach proposes grid fields of larger spacing to be the result of Moirè interference patterns. Blair, Welday, and Zhang (2007) argue in favour of the existence of a fundamental grid cell of very small spacing (coined "Theta cell") of an undetermined origin (suggested to integrate strong theta inputs from medial septum). Even if undetectable, a combination of two Theta cells differing only by a rotation or scaling of one of the patterns gives rise to the observed grid cell firing rate maps of arbitrary size and spacing via emerging Moirè interference patterns.

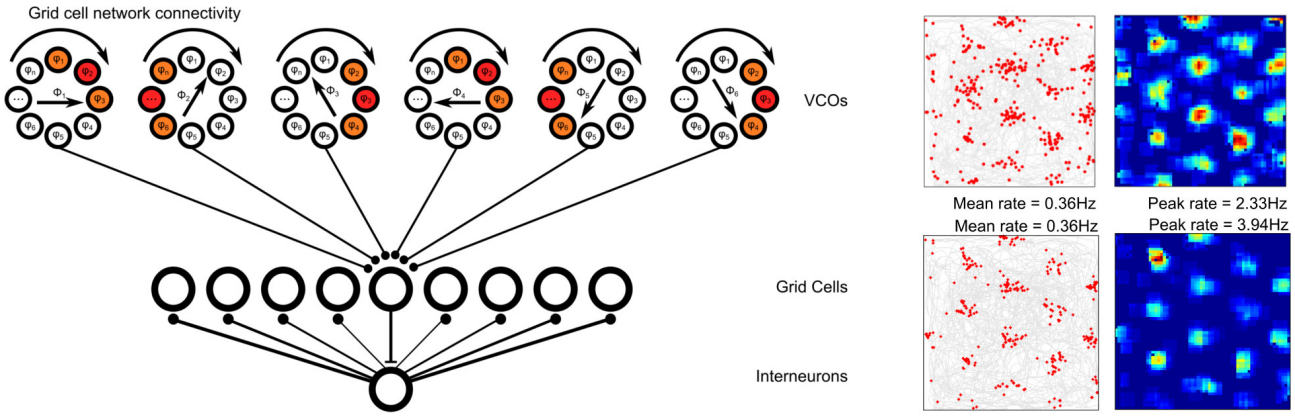


Figure 25: Hybrid OI/CAN model. *Left: model set-up. VCOs connect via inhibitory projections to a population of grid cells recurrently connected via interneurons. VCOs are arranged in independent ring attractors, with neurons in each ring sharing the same preferred direction but differing by a particular phase (warmer colour indicates higher activity). Grid cells inhibit each other with a strength profile that periodically follows grid field separation (simulated with a twisted torus topology). Right: spiking activity and firing rate maps of an example grid cell (top) and an example interneuron (bottom). Similar to CANs, interneurons are predicted to exhibit hexagonal periodicity. Modified with permission from Bush and Burgess 2014.*

What are experiments saying?

Continuous attractor and oscillatory interference models offer complementary descriptions of different aspects of grid cell activity. On the one hand CANs are able to answer questions regarding common properties found in a population of grid cells. On the other hand OIMs naturally give rise to the rhythmic temporal structure observed at the single cell level. While oscillatory activity can coexist within a CAN framework (see discussion above and Pastoll et al. 2013; Navratilova et al. 2012), the converse alternative has been only partially explored. Bush and Burgess (2014) examine the limitations of an OIM with additional CAN-like inhibitory connectivity (see Figure 25 for details on the model). Their investigations are focused on the effect of the additional connectivity on the quality of subthreshold membrane potential oscillations. In particular, they found the population entrainment offers spatial stability, robustness to noise, reduces

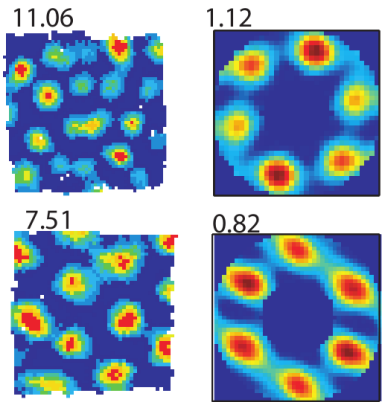


Figure 26: Turing defects in grid cells. *Examples of firing rates (left column, peak rate on top) and cropped autocorrelograms (right column, gridscore on top) for two cells in dMEC (rows) showing the appearance of lattice defects. Top row shows the splitting of the centre filed giving rise to a hepta/penta defect, while bottom row shows a penta defect. These defects are the signature of a Turing pattern formation process. Modified with permission from Krupic, Burgess, and O'Keefe 2012; selected examples from supplementary figure 2.*

previously strong theta activity, and gives rise to ramp-like depolarization profiles, better matching intracellular *in vivo* recordings (Schmidt-Hieber and Häusser 2013).

The main difference with a pure CAN model is that path integration is still performed through a phase code, and band-like VCO inputs must still interfere with a sixty degree separation among them. Can this mechanism coexist with a Turing-instability pattern formation process characteristic of CANs?, and does replacing the translation of the bump of activity in the neural sheet in favour of a phase code mechanism of path integration still retain the common population properties that are the signature of CANs?. These questions remain unanswered. In fact, on the absence of VCO inputs, the remaining inhibitory connectivity can support grid cell activity provided the network receives a tonic excitatory input. However, sustained hexagonal activity is accomplished by means of an intrinsic connectivity that is endowed with a twisted torus topology. This casts some doubt on the coexistence question, since sustaining a single bump of activity is in nature different than the emergence of hexagonality via Turing patterning (unless the twisted torus topology assumption is not a conceptual place-holder but an actual prediction of the physical connectivity of the network). Similarly, removing the concerted translation of activity in the neural sheet results in a lack of coherent activity at the population level, a feature supported by evidence of low dimensional attractors in experimental recordings (Yoon et al. 2013).

Robust variants of the oscillatory interference model require the existence of cells with periodic band-like fields. Krupic, Burgess, and O'Keefe (2012) recorded 351 cells from layers II and III of MEC and adjacent PaS, finding about a quarter of cells passed the standard arbitrary cut-off for grid cells. However, an additional half exhibited spatially stable multi-peaked fields, with strong biased preferences found in their two-dimensional Fourier components. The larger class of spatially periodic cells (SPCs) suggest it is improbable that grid cells' spatial maps are the result of precise combinations of band-like inputs, which would only produce perfect lattice arrangements [8].

On the contrary, the observed variety of multi-peaked maps lends its support to a two-dimensional pattern formation process. A good candidate is field self-organisation via Turing instabilities. The mechanism is currently exploited by continuous attractors and feed-forward models, albeit in very different ways. Independently of the model used, a hint that the outcome results from a Turing process are the typical pattern formation defects observed in the firing rate maps since the discovery of grid cells. Figure 26 shows the signature defects of Turing instabilities, namely splitting fields, penta/hepta

[8] A combination of a large number of band cells can in principle reproduce any given spatial map since they can serve as fundamental components of a Fourier-like decomposition. It would however require a large number of (unobserved) band cells spanning a wide spectrum of periodicities and orientations.

pairings, and a general proclivity of fields to appear close to the boundaries. A further clue is the influence on the patterns by particular shapes of the exploration boundary. In general, hexagonal fields tend to align in certain directions with respect to walls and local cues depending on the level of symmetry (or polarization) of the environment enclosure (Krupic et al. 2015; Stensola et al. 2015). Deformations of the pattern, including local changes of spacing, anisotropic sheering, and major distortions due to field dislocations all point towards boundary-induced defects of the Turing type. Defects arise in general from finite-size networks, noise in the centre-surround mechanism, the manner in which activity in the network is stabilised, and boundary effects.

Even though continuous attractor models implement the instability, they do so assuming an approximate Mexican-hat connectivity. An important consequence is that pattern formation happens on an abstract neural sheet, whose enclosure [9] would have a real impact on deformations of the pattern. More significant is the fact that CANs by themselves can not account for distortions of the hexagonal pattern due to physical environmental boundaries, a result that comes out naturally in feed-forward models.

[9] The actual geometry of the enclosure depends on the exact neuronal connectivity and can in principle have any arbitrary shape. However, for local networks with large number of neurons the defects are more strongly visible near the boundary.

An alternative: Why feed-forward models?

The entire class of feed-forward models rely on developing a Turing instability via a lateral inhibition mechanism (e.g. Mexican-hat dynamics). The early work by Kropff and Treves (2008) achieves a Mexican-hat temporal filtering of spatially modulated input by means of intermediate fatigue dynamics [10], which would implicitly describe cell adaptation. D'Albis and Kempter (2017) take this temporal filtering as their starting assumption to develop a mathematical analysis of the model, which was lacking in its original description.

The adaptation model requires the temporal filter kernel to have a Mexican-hat shape, otherwise pattern formation will not occur. It is therefore essential that actual firing rate adaptation, as recorded in experiments, can be modelled with a centre-surround filter. In experiments in vitro, when a constant input of a certain magnitude $h = h_0$ is suddenly turned on, the receiving cell almost instantaneously increases its firing rate from a previous steady baseline firing rate r_0 by a finite amount and undergoes firing rate adaptation immediately after, a process by which cell activity slowly decays and in some cases settle into a steady firing frequency in general larger than r_0. In contrast, the model describes a different picture. After the input onset, the cell increases its firing rate following the integral of the kernel $h_0 \int_0^t K$, which steadily grows until it reaches its zero and thereafter

[10] In the Kropff and Treves model, the update rules 3 and 4 for the fatigue variables can be rewritten as first-order differential equations

$$(1/b_2)\,\dot{r}_{inact} = h(t) - r_{inact}$$

$$(1/b_1)\,\dot{r}_{act} = h(t) - r_{inact} - r_{act},$$

with simple solutions given by

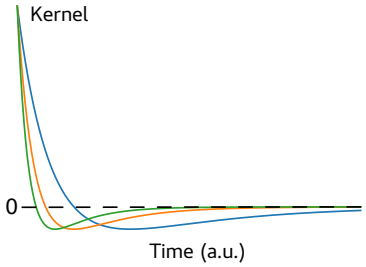
$$r_{inact} = b_2 e^{-b_2 t}\Theta(t) * h(t)$$

$$r_{act} = \frac{b_1}{b_1 - b_2}\left(b_1 e^{-b_1 t} - b_2 e^{-b_2 t}\right)\Theta(t) * h,$$

with the assumption that the input is turned on at time zero (h(t<o)=o). In their investigations, the ratio between the time constants $b_2 = b_1/3$ is kept constant, such that the temporal filter

$$r_{act} = \frac{3}{2} b_1 e^{-b_1 t}\left(1 - \frac{1}{3} e^{\frac{2}{3} b_1 t}\right)\Theta(t) * h,$$

depends only on a single free parameter. The previous equation has the form $r_{act} = K * h$, where the kernel



has a Mexican-hat shape as depicted for different values of b_1 (normalised amplitude). Since the inputs are temporally uncorrelated, the filter translates the dynamics into the learning of synaptic weights $J = \Gamma * J$, with Γ a Mexican-hat function in space.

decreases until it relaxes back to the starting value r_0. Typical measurements for the spacing of the hexagonal map and the natural exploratory speed of the animal put the decay time constant $1/b_1$ at around 100 ms and thus the zero crossing at about 150 ms (the actual values used in the models). In vitro recordings of EC layers II and III fit logarithmic and exponential adaptation time constants in that approximate range in response to strong step current inputs (Alonso and Klink 1993; Yoshida, Jochems, and Hasselmo 2013). However, firing frequency invariable and monotonically decreases in time as a result of the adaptation process. In particular, no increase is observed in time spans larger than the adaptation constants (e.g. 150 ms), with the initial rapid growth limited to undetectable ranges less than an order of magnitude below [11].

[11] There might however be an adaptation mechanism in vivo absent in in vitro recordings. No hints of such adaptation are apparent in the few in vivo experiments, but further systematic investigations are needed.

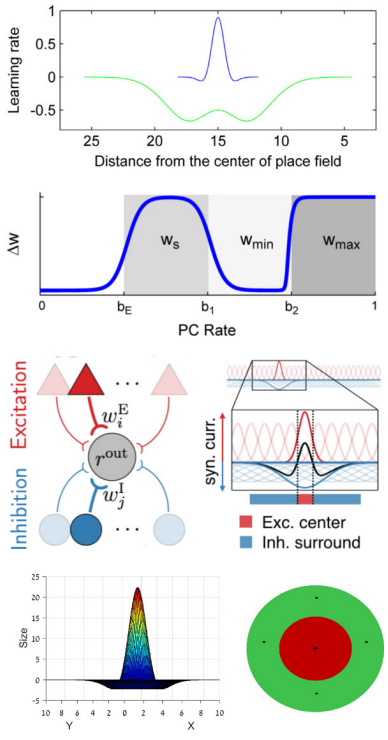


Figure 27: Spatial Mexican-hats in feed-forward models. *Hebbian plasticity must in the end implement a spatially tuned centre-surround learning rule. This can be accomplished by an activity-dependent transition from LTD to LTP (top), or alternatively by intermediate subpopulations of inhibitory and excitatory neurons, either in combination with novelty-induced activity modulation (middle-top) or by assuming broader tuning curves for the inhibitory interneurons (middle-bottom). As a straightforward way to analyse these models one can set the inputs from the start to have centre-surround spatial tuning (bottom). Modified with permission from Stepanyuk 2015; Castro and Aguiar 2014; Weber and Sprekeler 2018; Dordek et al. 2016.*

Other feed-forward models take a different approach to come up with centre-surround dynamics (see Figure 27). Stepanyuk (2015) starts from the assumption that there exists a consistent transition from LTD at low pre- and postsynaptic firing rates to LTP at higher rates in projections from hippocampus to EC (see Figure 27 top). Since place cells code space via their firing rates, this transition induces positive plastic changes close to the centre of the place field, and negatives changes otherwise (for fixed postsynaptic activity). Standard associative Hebbian plasticity thus translates the changes into a spatial Mexican-hat learning rule for the incoming weights. The model must be contrasted with realistic hippocampal data, where high firing rate variability at the population level and place field heterogeneity is commonplace. To preserve similar spatial regions undergoing different plastic changes, the presumed transition in plasticity needs to be established from cell to cell and depending in the overall cell activity (under an unspecified mechanism).

Castro and Aguiar (2014) take a step further and introduce a subpopulation of intermediate inhibitory neurons, such that place cells now communicate via excitatory and inhibitory pathways. An additional gain modulation mechanism is introduced, which emulates the observed expansion of place field size during exploration of novel environments. Together, the network architecture and activity modulation, transform a Hebbian mechanism into a spatially dependent centre-surround learning rule, with a centred disk of potentiation, followed by a ring of depression and a further ring of potentiation (equivalent to the action of a Mexican-hat in a pattern formation context; see Figure 27 middle-top). The learning rule suffers from the same downsides as the model above, although at least the suggested network architecture has been shown to exist between deep and superficial layers of EC. The prediction, however, that there exists interneurons with ring-like spatial firing (no firing in the centre) has

not been confirmed.

Weber and Sprekeler (2018) propose a similar solution with a simpler assumption. In their model, cells in the hippocampal formation receive input from subpopulations of spatially tuned excitatory and inhibitory cells. The major assumption is to consider the inhibitory population to have a broader spatial tuning compared to the excitatory cells. The combined effect turns simple Hebbian plasticity into the desired centre-surround learning kernel. The assumption requires however inputs to be spatially stable during the learning process. Even though interneurons in superficial layers of EC tend to be broadly tuned, they are highly interconnected to the grid cell population. As a result, it is unlikely they remain stable during periods of learning, given the dynamical interactions with the still developing grid cell patterns [12].

As a final example in the class, Dordek et al. (2016) examine the model from the perspective of principal component analysis. To update the weights, they use a simple self-normalizing Hebbian learning rule (Oja). In the absence of any additional constraints they find that the inputs themselves must be of the centre-surround type for the network to result in non-trivial spatial patterns. Regular, positive-only Gaussian-like firing rate inputs results in unremarkable spread activity. Furthermore, they are the first to explicitly remark on the effect of an essential constraint all models have used implicitly or just mentioned. A non-negativity constraint on the weights is indispensable for hexagonal patterns to be selected. The otherwise selected eigenmode reflects the symmetry given by the specific geometry of the enclosure, in their case they obtain a square lattice in a square enclosure. Even though a rather abstract description [13], the analysis makes explicit the need for centre-surround effective inputs and keeping the dynamics bounded from below.

[12] The possibility that stronger stabilizing inputs dominate interneuron firing is a real one. Strong projections from deeper layers of EC as well as subcortical inputs such as septoentorhinal projections have been shown to target interneurons directly.

[13] An alternative to negative firing rates is also proposed. Before feeding the inputs to the downstream cell, the activity is dynamically normalized by subtracting the (temporal) mean activity of each input cell. Biologically, this could be accomplished by some form of activity-dependent cell adaptation.

Evidence for feed-forward models

Despite these models' drawbacks, the appearance of typical Turing defects in grid cell spatial maps and additional experimental findings weight nonetheless in favour of feed-forward models. Bonnevie et al. (2013) showed that hippocampal back-projections are essential for stable grid cell firing. Muscimol inactivation of hippocampus completely silences spatial place cell activity 20 minutes after infusion. Entorhinal grid cells lost hexagonality soon after while retaining strong theta rhythms [14] (power dropped along with overall firing rates).

Wills et al. (2010) and Langston et al. (2010) characterize the developmental timeline of spatial representation in the hippocampal

[14] A finding that poses additional constraints on oscillatory interference models.

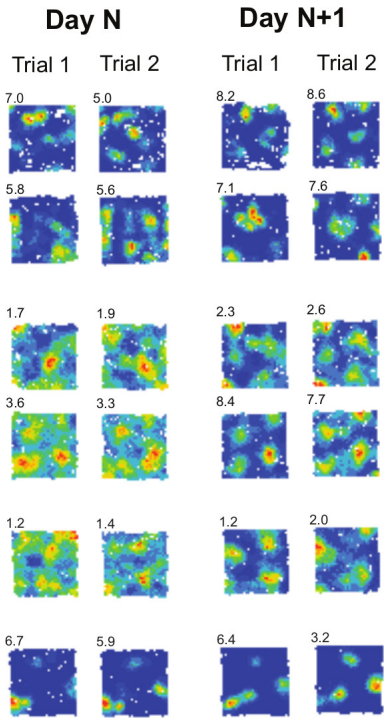


Figure 28: Development of the grid cell pattern. *Each row tracks a single cell over the span of one day until the first appearance of the hexagonal pattern. The first two cells were recorded on P19(20), cells 3 and 4 on P21(22), and cells 5 and 6 on P22(23) and P23(24) respectively. On each day two consecutive trials were recorded. Modified with permission from Wills, Barry, and Cacucci 2012.*

formation. Already in their first exploration of the environment after leaving the nest (around post-natal day 14), pups exhibit adultlike head-direction cells in pre- and parasubiculum, while place cells in area CA1 have close to adult levels of spatial information content, and high degrees of theta modulation and phase precession, however the number of place cells and the spatial stability of fields continue to increase until it reaches adult levels after around P30. In contrast, grid cells spatial maps show the slowest developmental times. Already from the outset MEC cells present rudimentary multipeak distributed activity, however only after around weaning age (P21) do the patterns improve in periodicity and spatial stability accompanied by and increasing number of theta modulated neurons and synchrony within the network. A more sudden change at around P28-30 brings the population of MEC cells close to adult levels. Importantly, developmental changes are correlated with age and not amount of spatial exploration, suggesting a degree of instructive hardwiring in the network that sets up the spatial representation system. Nonetheless, a baseline amount of exploration is unavoidable given the nature of the measurements, which could in fact be necessary for the development of functional properties in the system.

Wills, Barry, and Cacucci (2012) expand on these findings to show that the appearance of the hexagonal pattern is rather abrupt and is accompanied by adult-like spatial stability and to some extent greater spatial information. What is more, they identify six grid cells recorded over a day before robust hexagonality was first present. Figure 28 demonstrates once again the typical developmental line of a Turing pattern formation process. These recorded examples show how the initial spread of activity cluster into fields, and the subsequent rearrangement into a hexagonal lattice. Spatial exploration seems to be a necessary requisite to increase hexagonality, and times of the order of the duration of typical trials are in general enough to correct the pattern.

Even though individual grid cells develop hexagonal maps at different times, a significant fraction starts showing a more symmetrical arrangement around weaning age (post-natal day 21). Muessig et al. (2015) report that the sudden appearance of hexagonal patterns in MEC coincides with a change in place cell accuracy in area CA1. More precisely, CA1 spatial maps increase in stability and spatial information, while there is an overall shift in population encoding, from a biased distribution over-representing the boundaries to a more evenly spread representation of the enclosure. They hypothesize that recently developed hexagonal activity stabilizes CA1 spatial maps at the centre of the enclosure via intermediate projections to DG and CA3. While there is certainly a visible effect of MEC in hippocam-

pal areas, it is also worthwhile to consider the role of contributions during development in the reverse direction.

Donato et al. (2017) characterize the maturation timeline of cells in the hippocampal formation of mice. Maturation levels are associated with a measure of dendritic growth and stabilization of synaptic connectivity into a local network. Figure 29 shows the maturation timelines of principal neurons, interneurons, and synaptic density in different regions of the entorhinal-hippocampal network. In all cases the MEC-L2 subnetwork shows the first signs of adult levels of development. Stellate cells of MEC-L2 reach maturity as early as P14 (initial exploration age), and are shown to drive maturation of the whole circuit in an order resembling the flow of information among subnetworks. Notwithstanding, grid cells spatial maps are the latest to robustly appear in adult levels around P30. How can we reconcile the developmental timecourse of the network with theoretical descriptions of the emergence of hexagonality in grid cells?

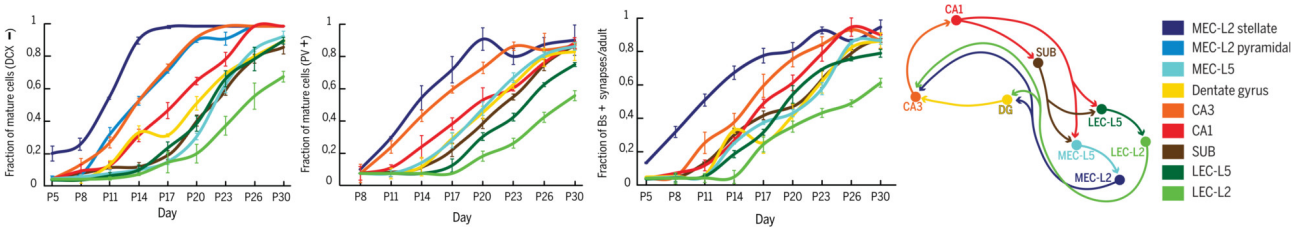


Figure 29: Maturation timecourse of the entorhino-hippocampal circuit. *The figure shows different fractional measures of maturity normalized to adult levels across all subnetworks during post-natal development. Left: overall fraction of mature cells (not expressing doublecortin). Middle: fraction of mature fast-spiking inhibitory cells (expressing parvalbumin). Right: density of synaptic punta in each local neuropil volume. Modified with permission from Donato et al. 2017.*

Inhibitory subpopulations in MEC-L2 and CA3 are close to adulthood around weaning age, presumably responsible for the first traces of periodicity and synchronization in grid cells, and the shift to a more evenly distributed population coverage of spatial representation in area CA1. Despite an almost mature MEC-L2 network, the most visible improvement of hexagonality at the population level coincides with an abrupt increase on synaptic density in MEC-L5, which together with higher fractions close to adulthood of MEC-L5 neurons, convey rich spatial information coming from now mature cells in CA1 and SUB into pyramidal and stellate cell populations in MEC-L2.

Altogether, the experimental evidence suggests a prominent role of feedback hippocampal pathways in setting robust hexagonal patterns. Yet, the incoming rich spatially modulated input does not tell a complete story. The underlying modulation by oscillatory rhythms (e.g. Theta) turns out to play an equally prominent role.

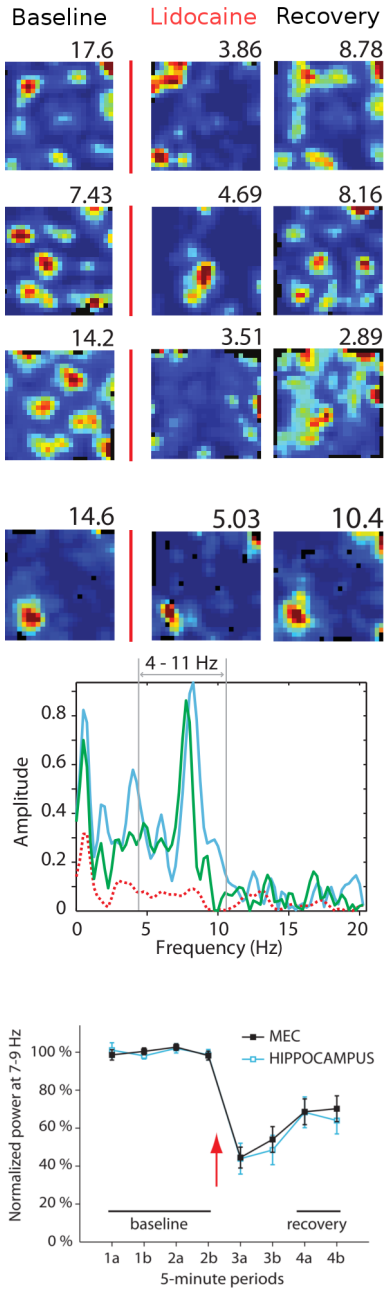


Figure 30: Lidocaine inactivation of septal input. *Hexagonality is severely affected shortly after inactivation of medial septum input (top). Other spatially rich cells are largely unaltered by the inactivation, even though rhythmic organization is disrupted (middle; spectrum of auto-correlation with blue:before, red:during, green:after). In general, a network-wide drop in theta power is observed all through the hippocampal formation (bottom). Modified with permission from Koenig et al. 2011.*

What our model brings to the table

Our model belongs to the general class of feed-forward models (Monsalve-Mercado and Leibold 2017). However, it stands out from all other models in that precise spike-timing relationships modulated by local rhythms bring about the Turing pattern formation process. In our model, a large ensemble of phase precessing spatially modulated inputs projects to a single entorhinal cell, whose synaptic connectivity is then set through spike-timing dependent plasticity. The learning window serves as a band-pass filter of the correlations among the input, with its shape being largely irrelevant for the emergence of hexagonality.

An ideal test of the model, which would in addition set it apart from other feed-forward models, would be to disrupt hippocampal spike-timing relationships, as orchestrated by the network oscillatory rhythms, while at the same time keeping their rich spatial information intact. Brandon et al. (2011) and Koenig et al. (2011) pharmacologically inactivated medial septum input to the hippocampal formation (see Figure 30). As a result, temporal organization within subnetworks is disrupted as evidenced by a drop in theta power in the local field potential and spike-train correlations, as well as a decrease in the quality of phase precession within MEC. Behavioural recordings show that during inactivation of MS most spatially modulated cells are functionally spared, with similar levels of spatial information before and after inactivation took place, including head direction, border, and spatially rich non-grid cells in MEC, and place cells in CA1 and CA3. Surprisingly, only grid cells' hexagonal patterns are affected by the loss of temporal coordination in population spiking activity, momentary losing hexagonality while still retaining strong spatial modulation. The experiments are far from being the ideal test for the model, on the one hand substantial theta power and phase precession remain after inactivation, although the observed significant drop could account for the lost of hexagonality. On the other hand, a strong excitatory input to the MEC and hippocampus has been suppressed (its major inhibitory projection targets mostly inhibitory interneurons in MEC), as evident by the overall decay in activity. In any case, the results highlight the importance of oscillatory activity in maintaining grid cells' hexagonality, in addition to a wealth of experimental reports linking the cells' temporal and spatial responses.

The main assumption of our model relies on the existence of a particular relationship between place cells' temporal and spatial activity on the population level. In a nutshell, cells whose place fields are close together are positively correlated in the theta band, while

the contrary is true for cells whose fields are further apart. It is neither a trivial statement nor has it been reported in the literature so far (but see Figure 32). We have mathematically derived this relationship taking the well-established phase precession phenomenon as our starting point. Our work points towards two important conjectures. First, in an open field exploratory task, and contrary to a linear track set-up, the spike-time cross-correlations of any two cells should be approximately symmetric for time lags within theta timescales. As a consequence, the oscillatory component of the cross-correlation (a band-pass filter around theta) should have a phase close to 0 or 180 degrees, thus respectively making the cells correlated or anti-correlated in the theta band. Second, pair cells belonging to the first category (positively correlated) are predicted to have place fields with a high degree of overlap (close together), while pairs in the second category (negatively correlated) would have lower -but still substantial- overlap between fields.

Regardless of the underlying mechanism responsible for establishing the relationship, it is possible to confirm directly from recorded hippocampal data whether our predictions have any degree of validity.

We examined a small set of simultaneously recorded hippocampal CA1 place cells (see Figure 31 for details). With a simple analysis, we found a propensity for correlation phases to be distributed around 0 and 180 degrees, supporting the claim that pair cells cross-correlations are prone to high extents of symmetry. A proxy measure for the overlap between place fields furthermore reveals a trend linking the quality of the correlation to the fields' overlap: The higher the overlap, the strongest the tendency for pairs of cells to be positively correlated in the theta band.

These results distinctly hint towards our model assumptions being plausible. However, the available data is insufficient to make a stronger definitive statement. For once, given that it is control data, the experiment was not optimally designed to confirm our predic-

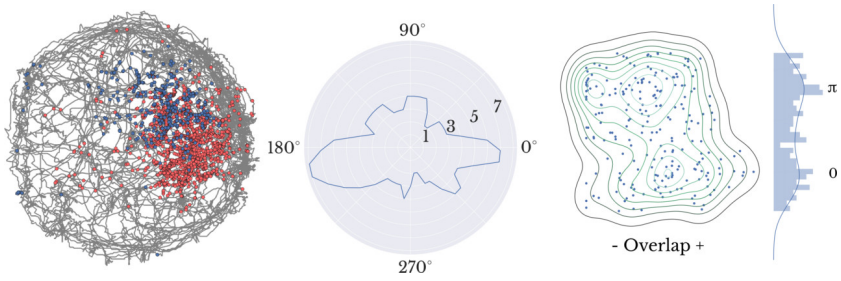


Figure 31: Phase correlations in the theta band. *Place cell recordings in a 1m diameter circular arena (control data from Schlesiger et al. 2015, left: spikes from two different cells) confirm that in the open field case the phases of pair-wise cross-correlations are clustered around 0 and 180 degrees (centre: phase histogram). Further analysis reveals a tendency for cells with higher overlap to have higher correlation in the theta band (right). The data includes 18 simultaneously recorded CA1 place cells. Recording sessions comprise six consecutive 15-20 minutes trials (with breaks in between) in a free foraging task. We then computed the cross-correlation of spike trains, filtered the signal for a range within 7-9 Hz, and obtained the phase of the oscillation from the Hilbert transform. A measure for the overlap was obtained from the logarithm of the interior product of place fields. Here, a place field was taken as the 2D histogram of the spike distribution with a 4 cm binning (no additional smoothing nor occupancy correction was performed).*

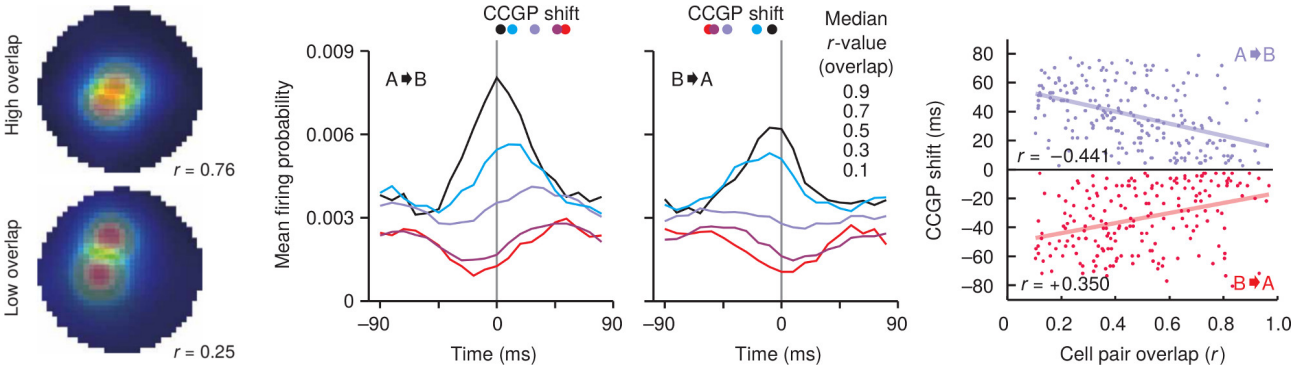


Figure 32: Phase correlations in the theta band: the 1D case. *During an open-field foraging task in a circular arena, 133 place cells were recorded from dorsal CA1 in six rats. The running path was then split into short traces to separate the data into runs going from one cell to another and in the reverse direction. For every cell pair, the spike-time cross-correlogram results from averaging over all traces, repeated for both categories or direction (from cell A to B and from cell B to A). Motivated by well-documented results in linear track experiments, the authors explore the relationship between the shift in the peak of the cross-correlogram and the amount of overlap of the two cells place fields. They observe that as the overlap diminishes the co-firing probability decreases and the peak moves away from zero. The direction of the shift follows the cell who fired last, the outcome being an approximate mirror quality when the categories are compared to each other. Modified with permission from Huxter et al. 2008.*

tions. For instance, many cells present poor inter-trial stability, one of the major contributors to the variance of the distribution in the form of extreme outliers. Restricting the analysis to a trial by trial examination increases the noise dramatically. In addition, low firing rates and almost vanishing overlap of many pairs are other major downsides of any dataset from which we would like to check our predictions. In general, a higher number of simultaneously recorded place cells in longer runs or more stable inter-trial performance would be necessary for a more definitive statement about the soundness of the assumptions.

Even though our predictions have not been explicitly investigated in the literature, a suggestive indication can be found in the work of Huxter et al. (2008) summarized in Figure 32. The path of the rats in an open field foraging task is subdivided into traces belonging to two categories. Traces that approximately cross the centres of two place fields in one direction form the first category, with the reverse orientation making up the second. The procedure is repeated for every pair of place fields. Averaging the spike-time cross-correlations over traces, trials, and rats, the position of the cross-correlogram peak (CCGP) is then contrasted to a measure of the place fields' overlap. Their findings reproduce what is known for linear track experiments, namely a shift on the cross-correlogram peak correlates with the amount of overlap.

However, an additional inference can be made from their results. Averaging (qualitatively by eye) the cross-correlograms corresponding to one direction with its reverse, it is possible to realise a distinct change in its phase for a critical amount of overlap. This is made more evident in Figure 32 rightmost plot. Averaging the two linear behaviours for the peaks' position results in an almost constant (around 0 ms) dependence with amount of overlap. Naturally, a 0

ms peak position is blind to the orientation of upward or downward phases, which is resolved by considering that the oscillations' period is around 120 ms, meaning that the linear average between ± 60 to ± 30 ms will correspond to a 180 degree phase, while the range ± 30 to 0 ms adds up to a 0 degree phase. The transition can be read off from the figure to happen at about 70% of their measure for the overlap.

Overall, the previous experimental reports strongly support the main assumption of our model. An open question of the model relates to the physiological mechanisms translating the input relationship into a form useful for the emergence of the pattern. Even though we raise the possibility that an appropriate learning window is able to fulfil this function, other mechanisms could be responsible as well, such as the resonant properties of stellate cells in EC layer II. In any case, additional experiments are required to test the essential role of input rhythms on the development of the firing patterns of grid cells.

Outlook

We propose an explanation for the emergence of grid cells firing patterns. In contrast to current models, the key element responsible for the emergence of hexagonality is already encoded in the firing dynamics of the input, a population of place cells upstream in the hippocampus. Just as other models, the basic physical process guiding the pattern formation is of the Turing type. The relevance of what particular mechanism induces the self-organisation of field maps is evident when we look at what specific predictions are possible to be confirmed via experiments. The idea is to find the simplest explanation possible that agrees the most with the already abundant wealth of experimental observations. Turing patterns are behind some of the most successful models by that measure, from continuous attractor networks to feed-forward models.

Recent experimental reports have brought forward attention to the effect that the environment has on the pattern. One of the strongest effects is the distortion of the pattern due to the degree of symmetry of the recording enclosure. These sets apart models based on continuous attractors from feed-forward ones, since in the former an abstract neural sheet receives the natural deformations caused by the Turing pattern formation process. In contrast, feed-forward models develop field regularity in physical space, in the sense that the self-organisation process feels the influence of geometrical boundaries and produces irregularities in the lattice such as penta/hepta defects, ellipticity, and shearing, which are also observed in recordings.

Further work might help corroborate some of the models predictions. An ideal test for the model involves perturbing the rhythmic modulation of place cells in order to lower the level of synchronisation at the population level, all the while retaining their spatial selectivity. It would be possible that such an experimental design might affect grid cells only during early post-natal development, perhaps with permanent impairment to the cells activity patterns. An alternative possibility suggests that also in adult animals a lower rhythmic modulation of activity might distort the pattern depending on the timescales of the learning mechanisms involved in their sustenance

that could induce smooth transitions out of hexagonality.

Recent reports moreover emphasize on the role that inhibition plays in shaping grid cell activity. Preliminary work considers a model where grid cells are coupled via a randomly distributed inhibitory network. We show that inhibition has a strong effect on the population patterns, implementing a self-correcting mechanism at the network level. Within this framework, it is now possible to explore the influence that the boundaries of the enclosure have on the geometrical properties of the hexagonal patterns. The self-correcting Turing formation process develops in a way that decreases defects in the patterns and increases hexagonality, a property that makes the process much more robust against several varieties of noise in the process. In addition, distortions due to the boundaries are natural to the system, reproducing the observed alignment of the lattice orientation to walls and local cues, and to some degree the ellipticity, shearing, and dependence of the grid spacing close to the boundaries (responsible for the reported barrel effect). Further analysis on the role of inhibition could address most findings concerning population properties of grid cell activity. Dynamical correlation, common geometrical properties within modules, and in general most properties related to population coherence, a hallmark of attractor networks, might be explained by the MEC recurrent collaterals. Moreover, the organisation of grid cells into discrete modules along the dorsoventral axis could come as a result of developmental hardwiring into clusters of inhibitory subnetworks or alternatively as a self-organisation process in which an extended network of grid cells would break into independent modules as a response to the gradient in size of the hippocampal inputs along the axis (or most likely a combination of both).

In our model, the shape of the learning window plays only a minor role and thus such framework might not be at all required. Plain activity-based Hebbian learning could be supplemented by an additional biological mechanism. A good candidate is the well-documented resonant sub-threshold dynamics of stellate cells in layer II (or other resonant types found as well in deeper layers). Their resonant frequency, coming from particular channel dynamics, agrees with the theta rhythms modulating the input activity of place cells, and is furthermore consistent all the way along the dorsoventral axis. Sub-threshold resonance might then take on the function of the learning window, acting as a filter on the theta range that extracts the Mexican-hat relation from the input correlations into the weight update rule.

The biological mechanisms behind the emergence of hexagonality are still under debate and awaiting experimental confirmation. How-

ever, it is likely that the fundamental process underlying grid fields regularity, largely supported by all recorded data so far, stands to be a Turing instability. Besides the emergence of grid fields, our model serves as a framework to study the dynamical spiking properties of grid cells, which have been neglected for the most part in Turing models.

Our work serves as a bridge between the temporal activity and spatial receptive fields of cells in the hippocampal formation. It takes one step forward towards the understanding of general coding principles and information processing mechanisms used by higher brain areas that deal with humans landmark ability to construct and work with abstract concepts and thoughts.

References

Alonso, A., and R. Klink. 1993. "Differential electroresponsiveness of stellate and pyramidal-like cells of medial entorhinal cortex layer II." *Journal of neurophysiology* 70, no. 1 (): 128–143. ISSN: 0022-3077.

Amaral, D. G., and M. P. Witter. 1989. "The three-dimensional organization of the hippocampal formation: a review of anatomical data." *Neuroscience* 31 (3): 571–591. ISSN: 0306-4522.

Amari, S.-i. 1977. "Dynamics of pattern formation in lateral-inhibition type neural fields". *Biological Cybernetics* 27 (2): 77–87. ISSN: 0340-1200. doi:10.1007/BF00337259.

Andersen, P., et al. 2006. *The Hippocampus Book*. Ed. by P. Andersen et al. Oxford University Press. ISBN: 9780195100273. doi:10.1093/acprof:oso/9780195100273.001.0001.

Barry, C., et al. 2007. "Experience-dependent rescaling of entorhinal grids". *Nature Neuroscience* 10, no. 6 (): 682–684. ISSN: 1097-6256. doi:10.1038/nn1905.

Barry, C., et al. 2012. "Grid cell firing patterns signal environmental novelty by expansion." *Proceedings of the National Academy of Sciences of the United States of America* 109, no. 43 (): 17687–17692. ISSN: 1091-6490. doi:10.1073/pnas.1209918109.

Barry, C., et al. 2006. "The boundary vector cell model of place cell firing and spatial memory." *Reviews in the neurosciences* 17 (1-2): 71–97. ISSN: 0334-1763.

Blair, H. T., A. C. Welday, and K. Zhang. 2007. "Scale-Invariant Memory Representations Emerge from Moiré Interference between Grid Fields That Produce Theta Oscillations: A Computational Model". *Journal of Neuroscience* 27 (12).

Bonnevie, T., et al. 2013. "Grid cells require excitatory drive from the hippocampus". *Nature Neuroscience* 16, no. 3 (): 309–317. ISSN: 1097-6256. doi:10.1038/nn.3311.

Brandon, M. P., et al. 2011. "Reduction of Theta Rhythm Dissociates Grid Cell Spatial Periodicity from Directional Tuning". *Science* 332, no. 6029 (): 595–599. ISSN: 0036-8075. doi:10.1126/science.1201652.

Brun, V. H., et al. 2002. "Place Cells and Place Recognition Maintained by Direct Entorhinal-Hippocampal Circuitry". *Science* 296, no. 5576 (): 2243–2246. ISSN: 00368075. doi:10.1126/science.1071089.

Burak, Y., and I. Fiete. 2006. "Do We Understand the Emergent Dynamics of Grid Cell Activity?" *Journal of Neuroscience* 26 (37).

Burak, Y., and I. R. Fiete. 2009. "Accurate path integration in continuous attractor network models of grid cells." *PLoS computational biology* 5, no. 2 (): e1000291. ISSN: 1553-7358. doi:10.1371/journal.pcbi.1000291.

Burgess, N. 2008. "Grid cells and theta as oscillatory interference: Theory and predictions". *Hippocampus* 18, no. 12 (): 1157–1174. ISSN: 10509631. doi:10.1002/hipo.20518.

Burgess, N., C. Barry, and J. O'Keefe. 2007. "An oscillatory interference model of grid cell firing". *Hippocampus* 17, no. 9 (): 801–812. ISSN: 10509631. doi:10.1002/hipo.20327.

Bush, D., and N. Burgess. 2014. "A hybrid oscillatory interference/continuous attractor network model of grid cell firing." *The Journal of neuroscience : the official journal of the Society for Neuroscience* 34, no. 14 (): 5065–5079. ISSN: 1529-2401. doi:10.1523/JNEUROSCI.4017-13.2014.

Castro, L., and P. Aguiar. 2014. "A feedforward model for the formation of a grid field where spatial information is provided solely from place cells". *Biological Cybernetics* 108, no. 2 (): 133–143. doi:10.1007/s00422-013-0581-3.

Conklin, J., and C. Eliasmith. 2005. "A Controlled Attractor Network Model of Path Integration in the Rat". *Journal of Computational Neuroscience* 18, no. 2 (): 183–203. ISSN: 0929-5313. doi:10.1007/s10827-005-6558-z.

Couey, J. J., et al. 2013. "Recurrent inhibitory circuitry as a mechanism for grid formation". *Nature Neuroscience* 16, no. 3 (): 318. doi:10.1038/nn.3310.

D'Albis, T., and R. Kempter. 2017. "A single-cell spiking model for the origin of grid-cell patterns". Ed. by V. Steuber. *PLOS Computational Biology* 13, no. 10 (): e1005782. doi:10.1371/journal.pcbi.1005782.

Dawitz, J., et al. 2017. "Transient Network Synchrony Of The Developing Medial Entorhinal Cortex". *bioRxiv*.

Donato, F., et al. 2017. "Stellate cells drive maturation of the entorhinal-hippocampal circuit." *Science (New York, N.Y.)* 355, no. 6330 (): eaai8178. ISSN: 1095-9203. doi:10.1126/science.aai8178.

Dordek, Y., et al. 2016. "Extracting grid cell characteristics from place cell inputs using non-negative principal component analysis". *eLife* 5 (): e10094. ISSN: 2050-084X. doi:10.7554/eLife.10094.

Dragoi, G., and G. Buzsáki. 2006. "Temporal encoding of place sequences by hippocampal cell assemblies." *Neuron* 50, no. 1 (): 145–157. doi:10.1016/j.neuron.2006.02.023.

Dunn, B., et al. 2017. "Grid cells show field-to-field variability and this explains the aperiodic response of inhibitory interneurons" ().

Egorov, A. V., and A. Draguhn. 2013. "Development of coherent neuronal activity patterns in mammalian cortical networks: Common principles and local hetereogeneity". *Mechanisms of Development* 130, **numbers** 6-8 (): 412–423. doi:10.1016/J.MOD.2012.09.006.

Egorov, A. V., et al. 2002. "Graded persistent activity in entorhinal cortex neurons". *Nature* 420, no. 6912 (): 173–178. ISSN: 0028-0836. doi:10.1038/nature01171.

Ermentrout, B., and D. H. (H. Terman. 2010. *Mathematical foundations of neuroscience*. 422. Springer. ISBN: 9780387877082.

Fuhs, M. C., and D. S. Touretzky. 2006. "A spin glass model of path integration in rat medial entorhinal cortex." *The Journal of neuroscience* 26, no. 16 (): 4266–4276. doi:10.1523/JNEUROSCI.4353-05.2006.

Fyhn, M., et al. 2007. "Hippocampal remapping and grid realignment in entorhinal cortex". *Nature* 446, no. 7132 (): 190–194. ISSN: 0028-0836. doi:10.1038/nature05601.

Fyhn, M., et al. 2004. "Spatial Representation in the Entorhinal Cortex". *Science* 305, no. 5688 (): 1258–1264. ISSN: 0036-8075. doi:10.1126/science.1099901.

Gierer, A., and H. Meinhardt. 1972. "A theory of biological pattern formation". *Kybernetik* 12, no. 1 (): 30–39. ISSN: 0340-1200. doi:10.1007/BF00289234.

Giocomo, L. M., M.-B. Moser, and E. I. Moser. 2011. "Computational models of grid cells." *Neuron* 71, no. 4 (): 589–603. ISSN: 1097-4199. doi:10.1016/j.neuron.2011.07.023.

Guanella, A., D. Kiper, and P. Verschure. 2007. "A model of grid cells based on a twisted torus topology". *International Journal of Neural Systems* 17, no. 04 (): 231–240. doi:10.1142/S0129065707001093.

Hafting, T., et al. 2005. "Microstructure of a spatial map in the entorhinal cortex." *Nature* 436, no. 7052 (): 801–806. ISSN: 1476-4687. doi:10.1038/nature03721.

Hartley, T., et al. 2014. "Space in the brain: how the hippocampal formation supports spatial cognition." *Philosophical transactions of the Royal Society of London. Series B, Biological sciences* 369, no. 1635 (): 20120510. ISSN: 1471-2970. doi:10.1098/rstb.2012.0510.

Hasselmo, M. E. 2008. "Grid cell mechanisms and function: Contributions of entorhinal persistent spiking and phase resetting". *Hippocampus* 18, no. 12 (): 1213–1229. ISSN: 10509631. doi:10.1002/hipo.20512.

Hopfield, J. J. 1982. "Neural networks and physical systems with emergent collective computational abilities". *Biophysics* 79:2554–2558.

– . 1984. "Neurons with graded response have collective computational properties like those of two-state neurons". *Proceedings of the National Academy of Sciences* 81, no. 10 (): 3088–3092.

Huxter, J. R., et al. 2008. "Theta phase-specific codes for two-dimensional position, trajectory and heading in the hippocampus". *Nature Neuroscience* 11, no. 5 (): 587–594. doi:10.1038/nn.2106.

Huxter, J., N. Burgess, and J. O'Keefe. 2003. "Independent rate and temporal coding in hippocampal pyramidal cells". *Nature* 425, no. 6960 (): 828–832. ISSN: 0028-0836. doi:10.1038/nature02058.

Jeewajee, A., et al. 2014. "Theta phase precession of grid and place cell firing in open environments." *Philosophical transactions of the Royal Society of London. Series B, Biological sciences* 369, no. 1635 (): 20120532. ISSN: 1471-2970. doi:10.1098/rstb.2012.0532.

Jung, M. W., S. I. Wiener, and B. L. McNaughton. 1994. "Comparison of spatial firing characteristics of units in dorsal and ventral hippocampus of the rat." *The Journal of neuroscience : the official journal of the Society for Neuroscience* 14, no. 12 (): 7347–7356. ISSN: 0270-6474.

Kjelstrup, K. B., et al. 2008. "Finite Scale of Spatial Representation in the Hippocampus". *Science* 321, no. 5885 (): 140–143. ISSN: 0036-8075. doi:10.1126/science.1157086.

Koenig, J., et al. 2011. "The Spatial Periodicity of Grid Cells Is Not Sustained During Reduced Theta Oscillations". *Science* 332, no. 6029 (): 592–595. ISSN: 0036-8075. doi:10.1126/science.1201685.

Kropff, E., and A. Treves. 2008. "The emergence of grid cells: Intelligent design or just adaptation?" *Hippocampus* 18 (12): 1256–1269. ISSN: 1098-1063. doi:10.1002/hipo.20520.

Kropff, E., et al. 2015. "Speed cells in the medial entorhinal cortex". *Nature* 523. ISSN: 0028-0836. doi:10.1038/nature14622.

Krupic, J., N. Burgess, and J. O'Keefe. 2012. "Neural Representations of Location Composed of Spatially Periodic Bands". *Science* 337, no. 6096 (): 853–857. ISSN: 0036-8075. doi:10.1126/science.1222403.

Krupic, J., et al. 2015. "Grid cell symmetry is shaped by environmental geometry". *Nature* 518, no. 7538 (): 232–235. ISSN: 0028-0836. doi:10.1038/nature14153.

Langston, R. F., et al. 2010. "Development of the Spatial Representation System in the Rat". *Science* 328, no. 5985 (): 1576–1580. ISSN: 0036-8075. doi:10.1126/science.1188210.

Leibold, C., and M. M. Monsalve-Mercado. 2016. "Asymmetry of neuronal combinatorial codes arises from minimizing synaptic weight change". *Neural Computation* 28 (8). ISSN: 1530888X. doi:10.1162/NECO{_}a{_}00854.

Leibold, C., and M. M. Monsalve-Mercado. 2017. "Traveling Theta Waves and the Hippocampal Phase Code". *Scientific Reports* 7 (1). doi:10.1038/s41598-017-08053-3.

Mathis, A., A. V. M. Herz, and M. Stemmler. 2012a. "Optimal Population Codes for Space: Grid Cells Outperform Place Cells". *Neural Computation* 24, no. 9 (): 2280–2317. ISSN: 0899-7667. doi:10.1162/NECO{_}a{_}00319.

Mathis, A., A. V. M. Herz, and M. B. Stemmler. 2012b. "Resolution of Nested Neuronal Representations Can Be Exponential in the Number of Neurons". *Physical Review Letters* 109, no. 1 (): 18103. ISSN: 0031-9007. doi:10.1103/PhysRevLett.109.018103.

Mathis, A., M. B. Stemmler, and A. V. M. Herz. 2015. "Probable nature of higher-dimensional symmetries underlying mammalian grid-cell activity patterns". *eLife* 4 (): e05979. ISSN: 2050-084X. doi:10.7554/eLife.05979.

Maurer, A. P., et al. 2005. "Self-motion and the origin of differential spatial scaling along the septo-temporal axis of the hippocampus". *Hippocampus* 15 (7): 841–852. ISSN: 1050-9631. doi:10.1002/hipo.20114.

McNaughton, B. L., et al. 1996. "Deciphering the hippocampal polyglot: the hippocampus as a path integration system." *Journal of Experimental Biology* 199 (1).

McNaughton, B. L., et al. 2006. "Path integration and the neural basis of the 'cognitive map'". *Nature Reviews Neuroscience* 7, no. 8 (): 663–678. ISSN: 1471-003X. doi:10.1038/nrn1932.

Miller, K. D., and F. Fumarola. 2012. "Mathematical Equivalence of Two Common Forms of Firing Rate Models of Neural Networks". *Neural Computation* 24, no. 1 (): 25–31. ISSN: 0899-7667. doi:10.1162/NECO{_}a{_}00221.

Moita, M. A. P., et al. 2003. "Hippocampal place cells acquire location-specific responses to the conditioned stimulus during auditory fear conditioning." *Neuron* 37, no. 3 (): 485–497. ISSN: 0896-6273.

Monsalve-Mercado, M. M., and C. Leibold. 2017. "Hippocampal Spike-Timing Correlations Lead to Hexagonal Grid Fields". *Physical Review Letters* 119, no. 3 (): 38101. ISSN: 0031-9007. doi:10.1103/PhysRevLett.119.038101.

Moser, E. I., et al. 2014. "Grid cells and cortical representation". *Nature Reviews Neuroscience* 15, no. 7 (): 466–481. ISSN: 1471-003X. doi:10.1038/nrn3766.

Muessig, L., et al. 2015. "A Developmental Switch in Place Cell Accuracy Coincides with Grid Cell Maturation." *Neuron* 86, no. 5 (): 1167–1173. ISSN: 1097-4199. doi:10.1016/j.neuron.2015.05.011.

Muller, R. U., and J. L. Kubie. 1987. "The effects of changes in the environment on the spatial firing of hippocampal complex-spike cells." *The Journal of neuroscience : the official journal of the Society for Neuroscience* 7, no. 7 (): 1951–1968. ISSN: 0270-6474.

Muller, R. U., J. L. Kubie, and J. B. Ranck. 1987. "Spatial firing patterns of hippocampal complex-spike cells in a fixed environment." *The Journal of neuroscience : the official journal of the Society for Neuroscience* 7, no. 7 (): 1935–1950. ISSN: 0270-6474.

Murray, J. D. 2004. *Mathematical Biology*. Ed. by J. D. Murray. Vol. 17. Interdisciplinary Applied Mathematics. New York, NY: Springer New York. ISBN: 978-0-387-95223-9. doi:10.1007/b98868.

Navratilova, Z., et al. 2012. "Phase precession and variable spatial scaling in a periodic attractor map model of medial entorhinal grid cells with realistic after-spike dynamics". *Hippocampus* 22, no. 4 (): 772–789. ISSN: 10509631. doi:10.1002/hipo.20939.

O'Keefe, J., and J. Dostrovsky. 1971. "The hippocampus as a spatial map. Preliminary evidence from unit activity in the freely-moving rat". *Brain Research* 34, no. 1 (): 171–175. ISSN: 00068993. doi:10.1016/0006-8993(71)90358-1.

O'Keefe, J., and L. Nadel. 1978. *The hippocampus as a cognitive map*. Clarendon Press.

O'Keefe, J. 2014. *Nobel Lecture: Spatial Cells in the Hippocampal Formation*.

– . 1976. “Place units in the hippocampus of the freely moving rat”. *Experimental Neurology* 51, no. 1 (): 78–109. doi:10.1016/0014-4886(76)90055-8.

O’Keefe, J., and N. Burgess. 2005. “Dual phase and rate coding in hippocampal place cells: theoretical significance and relationship to entorhinal grid cells.” *Hippocampus* 15 (7): 853–866. ISSN: 1050-9631. doi:10.1002/hipo.20115.

O’Keefe, J., and M. L. Recce. 1993. “Phase relationship between hippocampal place units and the EEG theta rhythm”. *Hippocampus* 3, no. 3 (): 317–330. ISSN: 1050-9631. doi:10.1002/hipo.450030307.

Pastoll, H., et al. 2013. “Feedback inhibition enables θ-nested γ oscillations and grid firing fields.” *Neuron* 77, no. 1 (): 141–154. ISSN: 1097-4199. doi:10.1016/j.neuron.2012.11.032.

Remme, M. W. H., M. Lengyel, and B. S. Gutkin. 2010. “Democracy-independence trade-off in oscillating dendrites and its implications for grid cells.” *Neuron* 66, no. 3 (): 429–437. doi:10.1016/j.neuron.2010.04.027.

– . 2009. “The Role of Ongoing Dendritic Oscillations in Single-Neuron Dynamics”. Ed. by K. J. Friston. *PLoS Computational Biology* 5, no. 9 (): e1000493. doi:10.1371/journal.pcbi.1000493.

Royer, S., et al. 2010. “Distinct Representations and Theta Dynamics in Dorsal and Ventral Hippocampus”. *Journal of Neuroscience* 30, no. 5 (): 1777–1787. ISSN: 0270-6474. doi:10.1523/JNEUROSCI.4681-09.2010.

Samsonovich, A., and B. L. McNaughton. 1997. “Path Integration and Cognitive Mapping in a Continuous Attractor Neural Network Model”. *Journal of Neuroscience* 17 (15).

Schlesiger, M. I., et al. 2015. “The medial entorhinal cortex is necessary for temporal organization of hippocampal neuronal activity”. *Nature Neuroscience* 18, no. 8 (): 1123–1132. ISSN: 1097-6256. doi:10.1038/nn.4056.

Schmidt-Hieber, C., and M. Häusser. 2013. “Cellular mechanisms of spatial navigation in the medial entorhinal cortex”. *Nature Neuroscience* 16, no. 3 (): 325–331. ISSN: 1097-6256. doi:10.1038/nn.3340.

Skaggs, W. E., et al. 1996. “Theta phase precession in hippocampal neuronal populations and the compression of temporal sequences”. *Hippocampus* 6, no. 2 (): 149–172. ISSN: 1050-9631. doi:10.1002/(SICI)1098-1063(1996)6:2<149::AID-HIPO6>3.0.CO;2-K.

Stemmler, M., A. Mathis, and A. V. M. Herz. 2015. "Connecting multiple spatial scales to decode the population activity of grid cells". *Science Advances* 1, no. 11 (): e1500816–e1500816. ISSN: 2375-2548. doi:10.1126/science.1500816.

Stensola, H., et al. 2012. "The entorhinal grid map is discretized". *Nature* 492, no. 7427 (): 72–78. ISSN: 0028-0836. doi:10.1038/nature11649.

Stensola, T., et al. 2015. "Shearing-induced asymmetry in entorhinal grid cells". *Nature* 518, no. 7538 (): 207–212. ISSN: 0028-0836. doi:10.1038/nature14151.

Stepanyuk, A. 2015. *Self-organization of grid fields under supervision of place cells in a neuron model with associative plasticity*, vol. 13. doi:10.1016/j.bica.2015.06.006.

Strange, B. A., et al. 2014. "Functional organization of the hippocampal longitudinal axis". *Nature Reviews Neuroscience* 15, no. 10 (): 655–669. ISSN: 1471-003X. doi:10.1038/nrn3785.

Taube, J. S., R. U. Muller, and J. B. Ranck. 1990. "Head-direction cells recorded from the postsubiculum in freely moving rats. I. Description and quantitative analysis." *The Journal of neuroscience : the official journal of the Society for Neuroscience* 10, no. 2 (): 420–435. ISSN: 0270-6474.

Thompson, L. T., and P. J. Best. 1990. "Long-term stability of the place-field activity of single units recorded from the dorsal hippocampus of freely behaving rats." *Brain research* 509, no. 2 (): 299–308. ISSN: 0006-8993.

Tolman, E. C. 1948. "Cognitive maps in rats and men." *Psychological Review* 55 (4): 189–208. ISSN: 1939-1471. doi:10.1037/h0061626.

Turing, A. M. 1952. "The Chemical Basis of Morphogenesis". *Philosophical Transactions of the Royal Society B: Biological Sciences* 237, no. 641 (): 37–72. ISSN: 0962-8436. doi:10.1098/rstb.1952.0012.

Weber, S. N., and H. Sprekeler. 2018. "Learning place cells, grid cells and invariances with excitatory and inhibitory plasticity". *eLife* 7 (): e34560. ISSN: 2050-084X. doi:10.7554/eLife.34560.

Welinder, P. E., Y. Burak, and I. R. Fiete. 2008. "Grid cells: The position code, neural network models of activity, and the problem of learning". *Hippocampus* 18, no. 12 (): 1283–1300. ISSN: 10509631. doi:10.1002/hipo.20519.

Widloski, J., and I. Fiete. 2014. "A Model of Grid Cell Development through Spatial Exploration and Spike Time-Dependent Plasticity". *Neuron* 83, no. 2 (): 481–495. doi:10.1016/j.neuron.2014.06.018.

Wills, T. J., et al. 2010. "Development of the Hippocampal Cognitive Map in Preweanling Rats". *Science* 328, no. 5985 (): 1573–1576. ISSN: 0036-8075. doi:10.1126/science.1188224.

Wills, T. J., C. Barry, and F. Cacucci. 2012. "The abrupt development of adult-like grid cell firing in the medial entorhinal cortex." *Frontiers in neural circuits* 6:21. ISSN: 1662-5110. doi:10.3389/fncir.2012.00021.

Witter, M. P., et al. 2017. "Architecture of the Entorhinal Cortex A Review of Entorhinal Anatomy in Rodents with Some Comparative Notes". *Frontiers in Systems Neuroscience* 11 (): 46. ISSN: 1662-5137. doi:10.3389/fnsys.2017.00046.

Yoon, K., et al. 2013. "Specific evidence of low-dimensional continuous attractor dynamics in grid cells". *Nature Neuroscience* 16, no. 8 (): 1077–1084. ISSN: 1097-6256. doi:10.1038/nn.3450.

Yoshida, M., A. Jochems, and M. E. Hasselmo. 2013. "Comparison of Properties of Medial Entorhinal Cortex Layer II Neurons in Two Anatomical Dimensions with and without Cholinergic Activation". Ed. by J. Brown. *PLoS ONE* 8, no. 9 (): e73904. ISSN: 1932-6203. doi:10.1371/journal.pone.0073904.

Zhang, K. 1996. "Representation of spatial orientation by the intrinsic dynamics of the head-direction cell ensemble: a theory". *Journal of Neuroscience* 16 (6).

Zilli, E. A., and M. E. Hasselmo. 2010. "Coupled noisy spiking neurons as velocity-controlled oscillators in a model of grid cell spatial firing." *The Journal of neuroscience : the official journal of the Society for Neuroscience* 30, no. 41 (): 13850–13860. ISSN: 1529-2401. doi:10.1523/JNEUROSCI.0547-10.2010.

Zilli, E. A., et al. 2009. "Evaluation of the Oscillatory Interference Model of Grid Cell Firing through Analysis and Measured Period Variance of Some Biological Oscillators". Ed. by B. S. Gutkin. *PLoS Computational Biology* 5, no. 11 (): e1000573. ISSN: 1553-7358. doi:10.1371/journal.pcbi.1000573.

Eidesstattliche Versicherung
Affidavit

Hiermit versichere ich an Eides statt, dass ich die vorliegende Dissertation *Space in the Brain: Of Learning and Representations* selbstständig angefertigt habe, mich außer der angegebenen keiner weiteren Hilfsmittel bedient und alle Erkenntnisse, die aus dem Schrifttum ganz oder annähernd übernommen sind, als solche kenntlich gemacht und nach ihrer Herkunft unter Bezeichnung der Fundstelle einzeln nachgewiesen habe.

I hereby confirm that the dissertation *Space in the Brain: Of Learning and Representations* is the result of my own work and that I have only used sources or materials listed and specified in the dissertation.

München, den 09.04.2018 Mauro Miguel Monsalve Mercado